INSTITUT FÜR DEUTSCHE SPRACHE

OPAL

Online publizierte Arbeiten zur Linguistik

ISSN 1860-9422

2/2011

Im Auftrag des Instituts für Deutsche Sprache
herausgegeben von Hardarik Blühdorn, Mechthild Elstermann und Annette Klosa

Technische Redaktion: Norbert Volz

Annette Klosa / Carolin Müller-Spitzer (Hg.)

Datenmodellierung für Internetwörterbücher

1. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“

Institut für Deutsche Sprache
Postfach 10 16 21
68016 Mannheim
opal@ids-mannheim.de

Mitglied der Leibniz Gemeinschaft

Inhalt

Einleitung

Annette Klosa klosa@ids-mannheim.de, Tel.: +49 621 1581-411
Carolin Müller-Spitzer mueller-spitzer@ids-mannheim.de, Tel.: +49 621 1581-429

Mittlerweile ist allgemein anerkannt, dass die Datenmodellierung für Wörterbücher in denjenigen lexikographischen Projekten, in denen aus einer Wörterbuchsubstanz verschiedene Wörterbücher in unterschiedlichen Medien hergestellt werden sollen, möglichst medienunabhängig erfolgen sollte. Geeignete Formate für eine solche medienunabhängige Datenmodellierung sind etwa XML-DTDs oder XML-Schemata, aber auch eine netzartige Modellierung in Relationen und Knoten. Neben wörterbuchspezifischen, maßgeschneiderten Modellierungen gibt es Richtlinien bzw. ‚Baukästen' für Standardmodellierungen, wie etwa das „Lexical Markup Framework for natural language processing (NLP) lexicons and machine-readable dictionaries (MRD)" (LMF, ISO-24613:2008)[1] oder die Richtlinien der „Text Encoding Initiative",[2] deren Einsatz für die Datenmodellierung bei Internetwörterbüchern zu diskutieren ist. Für elektronische Wörterbücher muss die Modellierung aber noch weiteren Anforderungen genügen: Hier bestimmen die erwünschten Zugriffsmöglichkeiten auf die Daten, wie diese modelliert werden müssen (vgl. Gloning/Welter 2001 und Müller-Spitzer 2005). Bei Internetwörterbüchern ist darüber hinaus bei der Modellierung zu berücksichtigen, dass eine flexible Präsentation der Suchergebnisse angestrebt werden soll, und zwar je nach Nutzergruppe oder Nutzersituation (vgl. Storrer 2001). Internetwörterbücher können solche flexiblen Zugriffs- und Präsentationsmöglichkeiten dann realisieren, wenn bei der Datenmodellierung die Funktionalitäten des Computers gleich mitgedacht wurden (vgl. u.a. de Schryver 2003).

Vor dem Hintergrund solcher Überlegungen fand am 5. und 6. Mai 2011 am Institut für Deutsche Sprache in Mannheim das erste Arbeitstreffen des wissenschaftlichen Netzwerks „Internetlexikografie" (gefördert von der Deutschen Forschungsgemeinschaft) statt.[3] Im Rahmen des Arbeitstreffens wurde die Modellierung verschiedener Internetwörterbücher in Konzeption und Realisierung vorgestellt und mit anderen, von konkreten Wörterbuchprojekten unabhängigen Modellierungsvorschlägen kontrastiert. Im Einzelnen wurde die Modellierung eines semantischen Netzes für lexikographische Anwendungen (am Beispiel der Duden-Ontologie; vgl. den Beitrag von Melina Alexa) präsentiert, die XML-Modellierung für ein Wörterbuchnetz (am Beispiel von OWID; vgl. den Beitrag von Carolin Müller-Spitzer), die TEI-basierte Modellierung von Retrodigitalisaten (am Beispiel des Trierer Wörterbuchnetzes; vgl. den Beitrag von Vera Hildenbrandt), eine FrameNet-basierte Modellierung (am Beispiel des Kicktionary; vgl. den Beitrag von Thomas Schmidt) sowie die Modellierung von Mehrwortverbindungen im DWDS (vgl. den Beitrag von Alexander Geyken). Außerdem wurde die Datenmodellierung bzw. Architektur für „pluri-monofunctional dictionaries" von Dennis Spohr vorgestellt.[4]

Dabei wurden Vor- und Nachteile des jeweiligen Vorgehens diskutiert, sodass sich zeigte, welche Formen der Datenmodellierung für welche Form von Internetwörterbüchern bzw. wörterbuchähnlichen Produkten am besten geeignet sind. Mit weiteren Fragen beschäftigten sich die Teilnehmer auch in einer Diskussionsrunde anhand verschiedener, für die Diskussion

[1] Vgl. http://www.lexicalmarkupframework.org/.
[2] Vgl. http://www.tei-c.org/index.xml.
[3] Zur Arbeit des Netzwerks „Internetlexikografie" vgl. http://www.internetlexikografie.de.
[4] Zu den Vorschlägen von Dennis Spohr für eine solche Modellierung vgl. detailliert Spohr (2011).

vorgegebener Fragen, deren Ergebnisse in den folgenden Abschnitten zusammengefasst werden.[5]

1. Welche Formate eignen sich am besten für die Modellierung von Internetwörterbüchern (DTDs, Schemata, Netze etc.)?

Hinsichtlich der Eignung verschiedener Formate für die Modellierung von Internetwörterbüchern war die einhellige Meinung der Diskussionsrunde: Komplexere Schemata sind im Hinblick auf ihre Mächtigkeit den DTDs überlegen. Außerdem können bestimmte funktionale Abhängigkeiten zwischen Attributen in ihnen expliziter formuliert werden, mögliche Verstöße gegen die formale Inhaltsstruktur sind somit besser zu kontrollieren. Dieser Vorteil ist jedoch weniger relevant speziell für die Präsentation von Internetwörterbüchern, vielmehr ist er für die Erstellung aller Arten von Wörterbüchern von Bedeutung. Wenn eine DTD für eine bestimmte Zielvorstellung nicht ausreichend ist, bietet sich daher der Einsatz von XML-Schemata an.

Zu bedenken ist auch, dass DTDs in Bezug auf die Definition von Datenformaten veraltet erscheinen, exakte Bestimmungen (wie etwa die Beschränkung auf ein Datumsformat) sind nicht möglich. Jedoch zeichnen sich DTDs durch den nicht unerheblichen Vorteil aus, dass sie für den Lexikographen lesbar sind. Man muss sich nicht auszugsweise über Probeartikel einen Überblick über die Modellierung verschaffen (den man auf diesem Weg zumindest bei einer komplexen Modellierung auch kaum bekommt), sondern kann die Modellierung in ihrer Gesamtheit in der DTD nachvollziehen; Entwürfe von DTDs können zwischen den Lexikographen und der Person, die sie modelliert, besprochen werden. Zudem sind DTDs für neue Wörterbuchprojekte automatisch in Schemata zu konvertieren. Die Beantwortung der Frage, ob Schemata oder DTDs herangezogen werden sollen, hängt somit vor allem davon ab, was in der Modellierung festgehalten werden soll, wie das lexikographische Team zusammengesetzt ist, wer die Modellierung lesen können soll und welche Rolle sie im Prozess der Wörterbucherstellung spielt.

Für bestimmte Ansätze sind DTDs als Grundlage möglich und ausreichend und insbesondere einfach zu handhaben. In einem Projekt wie *elexiko* beispielsweise ist die DTD sehr narrativ und auch für Uneingeweihte und Außenstehende nachzuvollziehen.[6] Andere Projekte, wie die Duden-Ontologie, benötigen eine netzbasierte Modellierung.[7] Die entscheidende Überlegung für die Wahl eines Modellierungsformates muss also – so der Tenor der Teilnehmer – sein: Welcher Formalismus kommt welchem Zweck zugute?

Es gilt auch zu bedenken, ob die modellierten und angehäuften Daten auf lange Sicht maschinell lesbar bleiben sollen und in welche Formate sie jeweils exportiert werden können. Generell kann festgehalten werden, dass das Verhältnis zwischen dem darzustellenden Inhalt und dem nötigen Aufwand bedacht werden muss, um die Entscheidung für die richtige Modellierungsmethode zu erleichtern. Schließlich muss man sich stets darüber klar sein, was und wie viel man investieren will, um die Daten langfristig kohärent zu halten. Neue Wörterbuchprojekte sind vor diesem Hintergrund auf jeden Fall gut beraten, die Frage der Datenmodellie-

[5] Die folgende Zusammenfassung beruht auf den Protokollen unserer Hilfskräfte Martin Loder, Bianca Pargner und Sandra Zimmermann, denen wir an dieser Stelle herzlich für ihre Unterstützung danken.

[6] Vgl. den Beitrag von Carolin Müller-Spitzer in diesem Band sowie Müller-Spitzer (2011).

[7] Vgl. den Beitrag von Melina Alexa in diesem Band.

rung ausreichend zu diskutieren, verschiedene Modellierungsansätze zu prüfen und genügend Zeit für die Umsetzung einzuplanen.

2. Was sind die Vor- und Nachteile projektspezifischer Modellierungen und allgemeiner Modellierungen wie TEI oder „Lexical Markup Framework“?

Entsprechend den unterschiedlichen Projekten, die die Diskussionsteilnehmer vertraten, waren sowohl Anwender von standardbasierten als auch von maßgeschneiderten Modellierungen vertreten. Allgemeine Übereinkunft bestand darin, dass eine vollkommene Austauschbarkeit von Daten zwischen verschiedenen elektronischen Wörterbüchern und Projekten auch bei einer Standardmodellierung nicht gewährleistet ist. Dafür ist die Bandbreite, die die Standardmodellierungen als Baukästen bieten, zu groß. Anders verhält es sich bei Retrodigitalisaten: Wenn in einer Institution (wie z.B. im Trierer Wörterbuchnetz) verschiedene Printwörterbücher in ein und demselben Modell für die elektronische Präsentation erfasst werden, bietet sich die Anwendung der TEI an.

Die TEI ist somit für die Modellierung neuer Wörterbücher eine gute Quelle, aus der man lernen kann. Ob allerdings die Anwendung einer TEI-konformen Modellierung generell als sinnvoller erachtet werden muss als eine genau auf die Bedürfnisse des Projekts angepasste Modellierung, wurde bezweifelt. Letztere kann in einem solchen Fall die bessere Alternative sein, auch weil ihre Anwendung möglicherweise weniger komplex und besser überschaubar ist. Entscheidend für diese Frage ist auch, wie das lexikographische Team zusammengesetzt ist. Ist z.B. überhaupt jemand vorhanden, der von Grund auf eine neue Modellierung entwickeln kann oder muss sich der- bzw. diejenige an Standards orientieren?

3. Ist bei Standardmodellierungen wirklich Austauschbarkeit der Daten gewährleistet?

Die Weitergabe und der Austausch von größeren Datenmengen kann für verschiedene Wörterbuchprojekte von gegenseitigem Nutzen sein, wie allgemein konstatiert wurde. Vollkommene Austauschbarkeit ist dabei natürlich nie gewährleistet. Doch ist das Minimieren des dafür nötigen Aufwandes sehr erwünscht und durch Standards ermöglicht, wobei sich allerdings auch kleinere, nicht standardmäßig modellierte Datenmengen, wenn sie denn klar definiert sind, prinzipiell ohne großen Aufwand in ein allgemeineres Standardformat migrieren und somit austauschen lassen. Jedenfalls sind in der TEI viele philologische Prinzipien und Richtlinien repräsentiert, auf die man gut zurückgreifen kann, wenn es um flexible Modelle für eine je individuelle Modellierung geht.

4. Wie können/müssen bei der Modellierung verschiedene Zugriffsmöglichkeiten berücksichtigt werden?

Wenn man ein Wörterbuch schreibt, weiß man nicht von Anfang an, was potentielle Nutzer alles wissen wollen. Die Frage, die in der Runde diskutiert wurde, war daher u.a., ob man wirklich eine vollkommen nutzerunabhängige Datenmodellierung aufbauen kann bzw. wo Restriktionen hierfür liegen. Zum Beispiel ist die Formulierung einer Bedeutungsparaphrase in den meisten Fällen nicht nutzerunabhängig, sondern auf eine bestimmte Zielgruppe zugeschnitten. Man sollte sich auch fragen, ob sehr komplexe und/oder spezifische Daten letztlich

tatsächlich genutzt werden, d.h., ob der Aufwand für eine möglichst nutzerunabhängige Modellierung gerechtfertigt ist.

Andererseits geht es bei der Digitalisierung von Daten zumindest in wissenschaftlichen Kontexten ja auch darum, eine Forschungsgrundlage für die spätere wissenschaftliche Beschäftigung mit einer Datenmenge zu schaffen, deren Ergebnisse und Methoden zunächst noch nicht abzusehen sind. Folglich sind Datenkomplexität und ein spezifisches Informationsangebot durchaus auch von linguistischer Seite erwünscht. Der wissenschaftlichen Lexikographie bleibt vor diesem Hintergrund deshalb vor allem der Auftrag, neue Arten der Datenmodellierung und des Informationsangebots auszuprobieren und auch neue, bisher unbekannte Arten des Zugriffs anzubieten. Eine Modellierung sollte möglichst ein Maximum des Darzustellenden vorsehen und anstreben. Die Internetlexikographie entwickelt sich somit – zumindest was die Ebene der Datenbasis angeht – in Richtung der lexikalischen Datenbank.

5. Hat die Art der lexikographischen Primärquellen (Belegarchiv, elektronisches Textkorpus) Einfluss auf die Wahl und Art der Modellierung?

Unterschiedliche Quellen erfordern selbstverständlich unterschiedliche Behandlungen; hier ist die Effizienzfrage entscheidend. Belege aus einem elektronischen Textkorpus sind (je nach Korpus in verschiedenem Maße) anderer Natur als solche aus dem lexikographischen Belegarchiv im Zettelkasten. Vor allem bei der Verlinkung auf die Belege im Internetwörterbuch ist dies relevant. Für die Modellierung von Internetwörterbüchern ist also nicht so sehr die Frage nach der Art der Primärquellen von Belang, sondern diejenige danach, wie zugegriffen bzw. verlinkt werden und was zur Darstellung kommen soll.

6. Welche Rolle spielt die Modellierung im lexikographischen Prozess von Internetwörterbüchern?

Mittlerweile sind sich sowohl die Verantwortlichen in Forschungseinrichtungen als auch kommerzielle Geldgeber darüber im Klaren, dass Internetlexikographie nicht ohne eine stärkere Gewichtung der technischen Aufgaben auskommt, was sich insbesondere am Beispiel der Datenmodellierung zeigt. Trotzdem ist es häufig noch schwierig, kompetente technische Mitarbeiter für die wissenschaftliche Lexikographie zu gewinnen, da oftmals nur geteilte oder befristete, zuweilen auch unterbezahlte Stellen für technisches Personal angeboten werden, das sich deswegen eher an die freie Wirtschaft bindet. Eine Veränderung dieser Situation wurde von allen Diskussionsteilnehmern als äußerst wünschenswert erachtet.

7. Literatur

Gloning, Thomas/Welter, Rüdiger (2001): Wortschatzarchitektur und elektronische Wörterbücher: Goethes Wortschatz und das Goethe-Wörterbuch. In: Lemberg, Ingrid/Schröder, Bernhard/Storrer, Angelika (Hg.): Chancen und Perspektiven computergestützter Lexikographie: Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Tübingen, S. 117-132.

Lexical Markup Framework for natural language processing (NLP) lexicons and machine-readable dictionaries (MRD). Internet: http://www.lexicalmarkupframework.org/. (Stand: Oktober 2011).

Müller-Spitzer, Carolin (2005): Die Modellierung lexikografischer Daten und ihre Rolle im lexikografischen Prozess. In: Haß, Ulrike (Hg.): Grundfragen der elektronischen Lexikographie. elexiko − das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York, S. 21-54. (Schriften des Instituts für Deutsche Sprache 12).

Müller-Spitzer, Carolin (2011): Der Einsatz einer maßgeschneiderten, feingranularen XML-Modellierung im lexikographischen Prozess. In: Klosa, Annette (Hg.): *elexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. Tübingen: Narr, S. 173-191. (Studien zur deutschen Sprache 55).

de Schryver, Gilles-Maurice (2003): Lexicographers' Dreams in the Electronic-Dictionary Age. In: International Journal of Lexicography 16/2, S. 143-199.

Spohr, Dennis (2011): A Multi-layer Architecture for „Pluri-monofunctional" Dictionaries. in: Fuertes-Olivera, Pedro A./Bergenholtz, Henning: E-Lexicography – The Internet, Digital Initiatives and Lexicography. London/New York, S. 103-120.

Storrer, Angelika (2001): Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: Lemberg, Ingrid/Schröder, Bernhard/Storrer, Angelika (Hg.): Chancen und Perspektiven computergestützter Lexikographie: Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. Max Niemeyer: Tübingen, S. 53-69.

Text Encoding Initiative. Internet: http://www.tei-c.org/index.xml. (Stand: Oktober 2011).

Wissenschaftliches Netzwerk „Internetlexikografie". Internet: http://www.internetlexikografie.de. (Stand: Oktober 2011).

Die dynamische Verknüpfung von Kollokationen mit Korpusbelegen und deren Repräsentation im DWDS-Wörterbuch

Alexander Geyken geyken@bbaw.de, Tel.: +49 30 20370-390

1. Einführung

Die Beschreibung von Kollokationen im DWDS-Wörterbuch (DWDS = Digitales Wörterbuch der deutschen Sprache) nutzt die Möglichkeiten des digitalen Mediums in mehrfacher Weise: erstens dadurch, dass die Extraktion der Kollokationen mittels statistischer Methoden nahezu vollständig korpusbasiert erfolgt. Dadurch ist stets eine transparente und nachprüfbare Rückbindung von der im Wörterbuch aufgeführten Zitierform der Kollokation zu allen Korpusbelegen für diese Kollokation möglich. Zweitens wird es durch die Trennung der Kollokation im Wörterbuch von ihren Korpusbeispielen möglich, die Verbindung der Kollokation im Wörterbuch zu den Kollokationsbelegen dynamisch zu halten. Angesichts wachsender Korpora ist dies eine wichtige Eigenschaft, da damit die Aktualität des Wörterbuchs auf Belegebene stets gewährleistet ist. Eine dritte Charakteristik der Kollokationsbeschreibung im DWDS ist die Entkopplung von Kollokation und Belegen im Redaktionssystem. Im Redaktionssystem wird nur die Zitierform der Kollokation angegeben, die Rückbindung auf die Korpusbelege erfolgt dann über automatisch berechnete Verweise auf die Korpusbelege. Schließlich ergeben sich über die dynamische Verknüpfung von Zitierform und Korpusbelegen auch keine Platzprobleme: nur die Zitierform wird im Wörterbuch festgehalten, die Kollokationsbelege liegen in einer eigenen lexikalischen Datenbank und können nach Bedarf vom Nutzer des Systems angefordert werden.

In diesem Beitrag soll zunächst der Hintergrund des DWDS-Wörterbuchs dargestellt werden. Im zweiten Abschnitt erfolgt eine kurze Charakterisierung des im DWDS-Wörterbuch verwendeten Kollokationsbegriffs. Dessen Einbettung in die Wörterbuchstruktur des DWDS-Wörterbuchs wird im dritten Abschnitt beschrieben. Das eigentliche digitale Herzstück der Kollokationsbeschreibung im DWDS-Wörterbuch ist das DWDS-Wortprofil, eine auf syntaktischer Analyse und statistischer Auswertung basierende automatische Kollokationsextraktion, deren Grundlagen und Qualität in Abschnitt 4 dargestellt werden. In Abschnitt 5 soll anhand einiger Beispiele illustriert werden, wie die Arbeitsteilung der automatischen Kollokationen und der lexikographischen Intuition in der täglichen lexikographischen Arbeit aussieht. Schließlich geben wir im letzten Abschnitt einen Ausblick auf die künftige Arbeit.

2. Hintergrund: Das Projekt „Digitales Wörterbuch der deutschen Sprache“ (DWDS)

Ziel des an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) beheimateten Projekts Digitales Wörterbuch der deutschen Sprache (DWDS) ist die Schaffung eines „Digitalen Lexikalischen Systems“ – eines umfassenden, jedem Benutzer über das Internet zugänglichen Informationssystems, das Auskunft über den deutschen Wortschatz in Vergangenheit und Gegenwart gibt. Das Projekt ist in drei Phasen von jeweils 6 Jahren zwischen 2007 und 2024 geplant (Klein/Geyken 2010). Während der ersten Phase werden drei an der BBAW und ihren Vorgängerinstitutionen erarbeitete Wörterbücher digital aufbereitet und in das Informationssystem des DWDS eingebunden: das historische Deutsche Wörterbuch von Jacob Grimm und Wilhelm Grimm, das Etymologische Wörterbuch des Deutschen (erarbeitet

unter Leitung von Wolfgang Pfeifer) und das zwischen 1961 und 1977 publizierte Wörterbuch der deutschen Gegenwartssprache ([WDG]). Ferner werden die in den Projekten DWDS und Deutsches Textarchiv erstellten Korpora in das DWDS-Informationssystem integriert und gemeinsam mit den Wörterbüchern abfragbar gemacht. Insbesondere sind dies das zeitlich und nach Textsorten ausgewogene DWDS-Kernkorpus (Geyken 2007, 100 Millionen Textwörter), das historische Referenzkorpus des Deutschen Textarchivs (ca. 50 Millionen Textwörter, www.deutschestextarchiv.de) und die vorwiegend aus elektronischen Zeitungsquellen stammenden DWDS-Ergänzungskorpora mit einer Größe von derzeit etwa 2,5 Milliarden laufenden Textwörtern. Alle Korpora werden fortlaufend erweitert. Ein Kernstück der ersten Projektphase besteht in der Formalisierung und Überführung der elektronischen Fassung des WDG in eine lexikalische Datenbank: das DWDS-Wörterbuch (Herold/Geyken 2008, Herold 2011).

Das DWDS-Wörterbuch unterscheidet sich vom WDG nicht nur durch die strengere Schemabeschreibung, sondern stellt auch eine moderate lexikographische Überarbeitung des WDG dar, insbesondere hinsichtlich Änderungen von ideologisch belasteten Wortartikeln und der Anpassung an die neue Rechtschreibung (Klein/Geyken 2010). Ziel des DWDS-Wörterbuchs ist die Erstellung eines auf wissenschaftlichen Prinzipien beruhenden synchronen Wörterbuchs mit gewissen historischen Anteilen (im Wesentlichen: Belegchronologie und etymologische Angaben bzw. Wortgeschichten). Das DWDS-Wörterbuch wird während der zweiten und dritten Projektphase in den Jahren 2013 bis 2024 bearbeitet werden. Mit Beginn der zweiten Phase wird sich die Anzahl der Projektmitarbeiter auf 10 Lexikographen erhöhen, die eine Wortstrecke von etwa 25.000 hochfrequenten Lemmata bearbeiten werden, die entweder zu neu waren, um in das WDG aufgenommen zu werden oder aus anderen Gründen keine Aufnahme in das WDG fanden. In der dritten Phase soll der Grundbestand des DWDS-Wörterbuchs auf allen Ebenen der Mikrostruktur überarbeitet werden. Im Folgenden wird nur ein kleiner Ausschnitt daraus beschrieben, derjenige der Kollokationen.

3. Kollokationen im DWDS-Wörterbuch

Als Kollokationen werden im DWDS-Wörterbuch Mehrwortverbindungen codiert, deren Gesamtbedeutung sich zwar aus der Bedeutung der Einzelwörter erschließen lässt, die aber dennoch keine beliebige Kombinatorik aufweisen, sondern in gewissem Maße als Muster im Sprachgedächtnis abgespeichert sind. Eine pragmatische, weil für die lexikographische Arbeit praktisch verwendbare Charakterisierung der Kollokationen basiert auf Hausmann. Kollokationen im hausmannschen Sinne sind „typische, spezifische Zweier/Dreier Beziehungen zwischen Wörtern“, die aus Basis und Kollokator bestehen (Hausmann 1984). Diese Kollokationen werden im DWDS-Wörterbuch als Kollokationen vom Typ1 bezeichnet. Beispiele hierfür sind (die Basis ist hier fettgedruckt, der Kollokator kursiv) *schütteres* ***Haar***, *heikles* ***Thema***, ***Unfall*** *bauen* (*bauen* in dieser Bedeutung ist nicht ohne *Unfall* denkbar…). Die Terminologie im DWDS weicht hier von Hausmann ab, der unter Kollokationstypen die möglichen syntaktischen Relationen zwischen Basis und Kollokator fasst. Auf diese wird in Abschnitt 5.1 näher eingegangen.

Kollokationen vom Typ1 stehen im Gegensatz zu den unspezifischen Wortverbindungen wie *Haus bauen* oder *Buch kaufen*. Zwischen diesen beiden Polen stehen Kollokationen vom Typ2. Dies sind Mehrwortverbindungen, bei denen es mehrere Möglichkeiten für Kollokate gibt, die jedoch nicht beliebig, sondern semantisch oder pragmatisch motiviert sind. Beispiele hierfür sind: *Ball abspielen*, *Ball zuwerfen*, *Recht anwenden*, *Recht brechen*, *Recht auf Mitbestimmung*, *Recht des Stärkeren*. Beiden Typen (Typ1 und Typ2) von Kollokationen ist ge-

mein, dass mit ihnen die „nicht erwartbaren“ Mitspieler (Kollokatoren) beschrieben werden sollen. Grenzfälle zwischen Typ1 und Typ2 können dann auftreten, wenn es zwei oder mehr Kollokatoren gibt, von denen eines besonders häufig auftritt.

Kollokationen werden im DWDS-Wörterbuch auch abgegrenzt von den Verbindungen, deren Gesamtbedeutung gar nicht oder nur teilweise aus den Einzelbedeutungen erschließbar ist. Diese werden in der Werkstattsprache des DWDS-Wörterbuchs als Phrasem codiert. Wir verwenden die Bezeichnung „Phrasem“ hier in einem weiteren Sinn als in der linguistischen Literatur üblich, wo es nur idiomatische Wendungen bezeichnet. Darüber hinaus werden Phraseme im DWDS-Wörterbuch auch zur Codierung von Grußformeln, Sprichwörtern, rhetorischen Fragen etc. verwendet.

Eine weitere wichtige Charakteristik der Kollokationsbeschreibung im DWDS-Wörterbuch besteht darin, dass die Kandidaten für Kollokationen grundsätzlich aus dem statistischen Wortprofil des DWDS bezogen werden sollen (s. Abschnitt 5). Die Mehrzahl der Kollokationen, die über die statistische korpusbasierte Wortprofilanalyse für die Neueinträge extrahiert werden, ist vom Typ2, also distributionell nicht spezifisch, aber usualisiert. Banale bzw. gänzlich unspezifische, aber dennoch statistisch signifikante Kookurrenzen muss der Lexikograph aussortieren. Mit der Verwendung des statistischen Wortprofils als Basis für die Kollokationsbeschreibung entfällt für den Lexikographen auch die Auswahl geeigneter Belege für eine Kollokation. Diese werden nicht explizit in das Wörterbuch geschrieben, sondern im Nachhinein über die automatische Verknüpfung mit dem Wortprofil hinzugefügt.

4. Modellierung der Kollokationen im DWDS-Wörterbuch und Arbeit mit dem Schema

4.1 Schemabeschreibung im DWDS-Wörterbuch

Die Grundidee der Schemabeschreibung des DWDS-Wörterbuchs besteht in der konsistenten Auszeichnung auf allen Strukturebenen des Wörterbuchs, angefangen bei Formangaben und grammatischen Angaben, über Bedeutungsangaben, pragmatische Markierungen, Beispiele und Zusätze sowie Belege, bis hin zu den Verweisstrukturen. Das DWDS-Wörterbuchschema besteht aus ca. 50 verschiedenen Beschreibungselementen auf Element-Ebene sowie festen Wertelisten (beispielsweise bei grammatischen Angaben wie Genus oder Angaben zu Sach- oder Fachgebieten). Die Schema-Beschreibung des DWDS-Wörterbuchs liegt in Form von RELAX-NG und Schematron-Regeln vor und ist mit den Richtlinien der TEI P5 kompatibel. Die elektronische Fassung des DWDS-Wörterbuchs ist in diesem Schema validierbar (Herold 2011).

Für die anstehende Arbeit der Lexikographen wurde ein reduziertes Schema, eine sogenannte Werkstattsprache entworfen, die aber automatisch in das DWDS-Wörterbuchschema konvertierbar ist. Die Lexikographen erarbeiten die Artikel innerhalb eines lexikographischen Redaktionssystems.[8] Dieses besteht im Wesentlichen aus der lexikographischen Rechercheumgebung der DWDS-Website und einem DWDS-Framework, welches für die Autorenumgebung des oXygen-Editors entwickelt wurde. Die von dem Lexikographenteam erstellten Artikel werden zentral über das in eine graphische Benutzeroberfläche von oXygen integrierte Versionierungssystem Apache Subversion (SVN) verwaltet.

[8] http://www.dwds.de/projekt/lexarbeitsplatz/.

4.2 Modellierung der Kollokationen im XML-Wörterbuchschema

Die Kollokationen sind im DWDS-Wörterbuchschema in der Werkstattsprache innerhalb der Lesartenbeschreibung codiert (vgl. Abb. 1). Der Vorzug der Werkstattsprache besteht für die praktische lexikographische Arbeit darin, dass Mikrostrukturen, die im TEI-Schema eingebettet codiert sind, beispielsweise durch ein Attribut-Wert-Paar, in der Werkstattsprache explizit als Element codiert und somit im XML-Editor einfacher handhabbar sind. Ein Beispiel hierfür ist das Element Konstruktionsmuster in Abbildung 1, welches in einem TEI-P5-Schema als *cit type="pattern"* codiert werden müsste. Wenn man dann zusätzlich noch weitere Typisierungen, beispielsweise nach syntaktischer Klasse hinzufügen möchte, wird der Redaktionsprozess im XML-Editor schnell unübersichtlich.

Das Schema-Fragment der Lesartenbeschreibung ist rekursiv definiert und ermöglicht beliebig viele Einbettungen von Unterlesarten. Auf jeder Lesartenebene können Formangaben, die Syntagmatik, die Diasystematik, Frequenzangaben und Verweise beschrieben werden. Diese Informationen werden der Definition der Lesart vorangestellt. Auf die Definition wiederum folgen die Verwendungen, in denen Phraseme, Kollokationen (vom Typ1 oder Typ2, notiert als <Kollokation1> bzw. <Kollokation2>), Kompetenzbeispiele oder Korpusbelege beschrieben werden.

```
Lesart= element Lesart {
        Formangabe *
        , Konstruktionsmuster ?
        , Diasystematik *
        , Frequenzangabe ?
        , Verweise *
        , Definition ?
        , Verwendungen *
        , Lesart *
}

Verwendungen = element Verwendungen {
        Phrasem *
        &Kollokation1 *
        & Kollokation2 *
        & Beleg *
        & Kompetenzbeispiel *
}

Kollokation1 = element Kollokation1 {
        , attribute type { 'ATTR' | 'CJ' | 'OBJA' ... }
        , Zitierform
        , Paraphrase ?
        , Diasystematik?
        , Verweise
}
```

Abb. 1: Lesartenbeschreibung im DWDS-Werkstattschema

Eine Kollokation wird in ein <Kollokation1>- oder ein <Kollokation2>-Element eingeschlossen. Die Kollokation umfasst die Basis (das Stichwort, unter dem die Kollokation aufgeführt

wird) und den Kollokator. Diese werden nicht näher in der XML-Struktur unterschieden, sondern als Text in das Element <Zitierform> eingeschlossen (vgl. die Beispiele in Abbildung 2).

```
<Kollokation1 type= "OBJ">
        <Zitierform>Termin einhalten</Zitierform>
</Kollokation1>
<Kollokation1 type= "MOD">
        <Zitierform>sündhaft teuer</Zitierform>
</Kollokation1>
<Kollokation2 type= "OBJA">
        <Zitierform>Ball abspielen</Zitierform>
</Kollokation2>
```

Abb. 2: Beispiele für die XML-Struktur der Kollokationen

Falls es für ein Stichwort mehrere Kollokationen gibt, werden diese zunächst nach dem Kollokationstyp (Typ 1 und Typ 2), zweitens nach syntaktischen, drittens nach semantischen und schließlich nach pragmatischen Kriterien gruppiert. Die syntaktischen Gruppierungen werden im Werkstattschema über den Relationsnamen referenziert, der auf dem DWDS-Wortprofil (vgl. Abschnitt 5) basiert und damit die automatische Identifizierung der Korpusbelege über den Schlüssel im DWDS-Wortprofil ermöglicht. Die folgenden drei Beispiele illustrieren das Zusammenspiel dieser Kriterien.

Im ersten Beispiel (*Jeans*, vgl. Abbildung 3) findet eine Mischung von syntaktischen und semantischen Gruppierungen statt. Die syntaktischen Gruppierungen werden explizit typisiert. Dies geschieht durch die syntaktische Relation ATTR, die im Wortprofil die Adjektiv-Nomen-Relation bezeichnet. Die semantische Gruppierung wird hingegen nur über Kommentare (die in oXygen als „processing instructions“ codiert werden) festgehalten, da die semantischen Klassen zu offen sind, um sie in das enge Korsett einer geschlossenen Wertemenge zu pressen. Im vorliegenden Fall bezieht sich die semantische Gruppierung auf die Kriterien der Farblichkeit in der ersten Gruppe und die stoffliche Qualität in der zweiten Gruppe.

```
<Kollokation2 type= "ATTR">
        <Zitierform>helle Jeans</Zitierform>
        <Zitierform>dunkle Jeans</Zitierform>
        …
        <Zitierform>weiße Jeans</Zitierform>
</Kollokation2>
<Kollokation2 type= "ATTR">
        <Zitierform>abgetragene Jeans</Zitierform>
        <Zitierform>ausgebeulte Jeans</Zitierform>
        …
        <Zitierform>zerschlissene Jeans</Zitierform>
</Kollokation2>
```

Abb. 3: Kollokationen von *Jeans*

Im zweiten Beispiel (*Allergie*, vgl. Abbildung 4) werden drei Kollokationsgruppen gebildet, die sowohl syntaktisch als auch semantisch unterschiedlich sind: Gruppe 1 bezeichnet eine Substantiv-Koordination (Relation CJ), die die semantische Relation von Begriff/Oberbegriff zum Ausdruck bringt, Gruppe zwei die Eigenschaft (ATTR) und die dritte Gruppe bezeichnet schließlich eine inchoative Handlung in Form einer Nomen-Verb-Relation (OBJA). Codiert werden diese Kollokationen im DWDS-Wörterbuch wie folgt:

```
<Kollokation1 type= "CJ">
        <Zitierform>Allergien und Unverträglichkeiten</Zitierform>
</Kollokation1>
<Kollokation2 type= "ATTR">
        <Zitierform>eine heftige Allergie</Zitierform>
          <Zitierform>eine schwere Allergie</Zitierform>
          <Zitierform>eine starke Allergie</Zitierform>
</Kollokation2>
<Kollokation2 type= "OBJA">>
        <Zitierform>eine Allergie auslösen</Zitierform>
</Kollokation2>
```

Abb. 4: Kollokationen von *Allergie*

Im dritten Beispiel (*Handy*, vgl. Abbildung 5) werden syntaktisch gleiche Relationen (SUBJ) nach ihrer stilistischen Färbung gruppiert. Der Skopus des Elements Stilebene umfasst dabei alle Zitierformen des übergeordneten Kollokationsblocks. Diese Art der Codierung stellt ein Zugeständnis an die Anforderungen der täglichen Arbeit im Artikelredaktionssystem dar, bei der die Diasystematik einfach und einheitlich über alle Elemente zugreifbar sein muss. Bei der Konvertierung der Werkstattsprache in das DWDS-Wörterbuchschema werden diese „seriellen“ Abhängigkeiten wieder in hierarchische umgewandelt.

```
<Kollokation2 type= "SUBJ">
        <Zitierform>das Handy piepst</Zitierform>
        <Zitierform>das Handy bimmelt</Zitierform>
        <Diasystematik>
                <Stilebene>umgangssprachlich</Stilebene>
        <Diasystematik>
</Kollokation2>
<Kollokation2 type= "SUBJ">
        <Zitierform>das Handy klingelt</Zitierform>
        <Zitierform>das Handy vibriert</Zitierform>
</Kollokation2>
```

Abb. 5: Kollokationen von *Handy*

Die Gruppierungen lassen sich für die Präsentationsebene entsprechend ihren Typisierungen anzeigen oder nach alphabetischer Reihenfolge oder nach statistischer Signifikanz sortieren. Letztere wird über den Abgleich mit der Datenbank des DWDS-Wortprofils realisiert, die im folgenden Abschnitt beschrieben wird.

5. „Assistierte“ Kollokationsextraktion

5.1 Das DWDS-Wortprofil

Das DWDS-Wortprofil ist das Ergebnis einer automatischen syntaktischen und statistischen Analyse sehr großer Korpora. Es liefert einen kompakten Überblick über die statistisch signifikanten syntagmatischen Beziehungen eines Wortes mit anderen Wörtern. Beispiele dieser sogenannten syntaktischen Relationen sind Attribut-Nomen-Verbindungen wie *schöne Bescherung* oder Verb-Objekt-Beziehungen wie *Flasche entkorken*. Die Darstellung der Relationen erfolgt in Form einer Schlagwortwolke oder in Tabellenform. Das DWDS-Wortprofil beruht auf einer syntaktischen Voranalyse der Korpusdaten durch den Shallow Parser Syncop

(SYNtactic COnstraint Parsing, vgl. Didakowski 2007). Die Berechnung des DWDS-Wortprofils selbst erfolgt in drei Etappen: Festlegung der zu extrahierenden syntaktischen Relationstypen, Extraktion der Relationen mittels einer automatischen syntaktischen Analyse und Bewertung der statistischen Signifikanz der extrahierten Relationen. Die Methodik des Wortprofils ist anderweitig ausführlich beschrieben (Geyken et al. 2009). An dieser Stelle beschränken wir uns aus Platzgründen auf die praktischen Ergebnisse des Wortprofils.

Der derzeitige Prototyp des DWDS-Wortprofils (Wortprofil_2010) ist unter www.dwds.de abfragbar. Er beruht auf einer Mischung eines Referenz- und eines Zeitungskorpus, des DWDS-Kernkorpus und des ZEIT-Archivs (1946–2009), und hat eine Gesamtgröße von 500 Millionen laufenden Textwörtern. Aus dem Korpus wurden etwa 90.000 Lemmata mit 2.000.000 Relationen extrahiert. Anhand eines Beispiels sollen die verschiedenen, vom DWDS-Wortprofil extrahierten Informationen verdeutlicht werden. Beispielsweise werden für das Stichwort *Feindbild* im DWDS-Wortprofil 32 verschiedene syntaktische Relationen mit insgesamt 384 Vorkommen extrahiert. Diese werden in Form einer Schlagwortwolke dargestellt (vgl. Abbildung 6). Die Voraussetzung für die Aufnahme einer Relation in das Wortprofil ist, dass dafür wenigstens vier Belege im Korpus vorkommen. Damit soll verhindert werden, dass okkasionelle Verbindungen fälschlicherweise in das Wortprofil aufgenommen werden.

Die syntaktisch relevanten Nachbarn von *Feindbild* sind in den folgenden syntaktischen Relationstypen zu finden:

- Adjektiv-Nomen (Etikett: ATTR): altes, antibolschewistisches, äußeres, gemeinsames, gepflegtes, ideales, ideologisches, intaktes, klares, klassisches, linkes, neues, primitives, richtiges, schlichtes, überkommenes, verblasstes, westliches
- Nomen-Nomen (im Genitiv) (GMOD): Abbau, Verlust
- Nomen-Koordination-Nomen (CJ): Vorurteil
- Nomen -Verb (SUBJ):bleiben, stimmen, verblassen
- Nomen -Verb (OBJA): abbauen, aufbauen, brauchen, nehmen, schaffen
- Verb-Präposition-Nomen (V_PP): auskommen ohne, taugen als

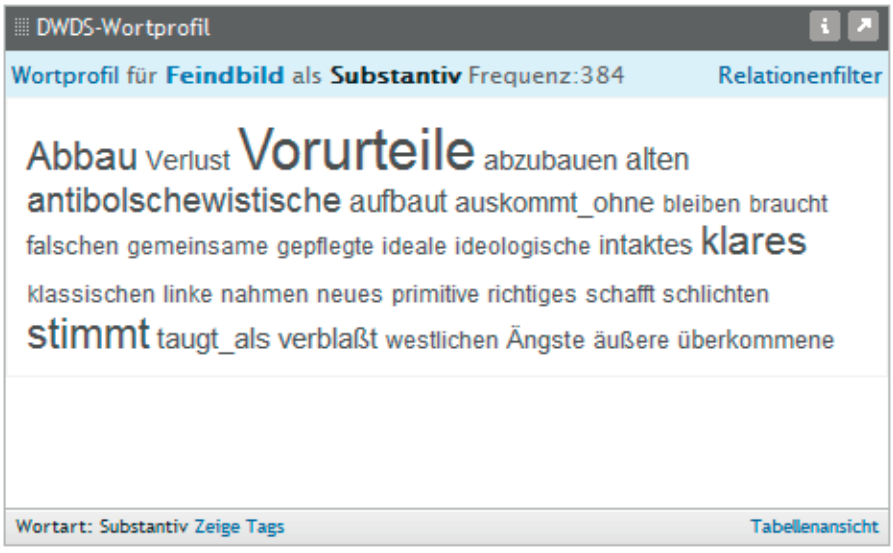


Abb. 6: DWDS-Wortprofil für *Feindbild*

Die Relationstypen lassen sich über das Wortprofil-Fenster ansteuern, indem man den Relationenfilter anklickt (vgl. Abbildungen 6 und 7). Es werden dann die syntaktischen Relationstypen aufgeklappt (vgl. Abbildung 7). Klickt man auf einen der Relationstypen, beispielsweise auf „Attribut“, erhält man alle Wortformen, die in einer Attributrelation (Adjektiv-Nomen) zum Wort *Feindbild* stehen. Diese Filter können bei hochfrequenten Wortprofilen sehr nützlich sein. Beispielsweise hat das bereits weiter oben erwähnte Substantiv *Haar* 13509 Relationen (davon 532 verschiedene) im DWDS-Wortprofil. Eine Darstellung als Schlagwortwolke wäre hier sehr unübersichtlich. Durch das Filtern nach einzelnen Relationstypen hingegen erhält man homogene Listen und handhabbare Größen.

Abb. 7: DWDS-Wortprofil für *Feindbild* – Relationstyp „Attribut“

Ein wesentlicher Mehrwert des Wortprofils besteht darin, dass alle extrahierten Relationen stets mit den dazugehörigen Satzkontexten im Korpus verknüpft sind und somit einen Überblick über den Verwendungszeitraum und die semantischen und pragmatischen Kontexte ermöglichen, in denen die syntaktische Relation verwendet wird. Klickt mal beispielsweise in Abbildung 6 auf die Verb-Verbindung *auskommen_ohne*, gelangt man zu den in Abbildung 8 gezeigten insgesamt vier Satzkontexten, die vom Analysesystem aus dem Korpus extrahiert wurden.

auskommen_ohne:
1 1993-10-22 | Zeitung:ZEIT Hamburg 1993; 206 S., 26, -DM fast ganz ohne dieses Feindbild auskommt:
2 1991-09-26 | Zeitung:ZEIT Hysterie und Haß Doch'die Unfähigkeit des einstigen Freiheitshelden gegen Kreml und Kommunismus, der jetzt ohne Feindbilder nicht mehr auskommt, der über den Weltmarkt nichts, aber über die Weltverschwörung gegen Georgien alles weiß, hat Hysterie und Haß gesät.
3 1989-10-20 | Zeitung:ZEIT „Das Land kam vorübergehend ohne Feindbilder aus ", heißt es im Begleitbuch.
4 1972-07-28 | Zeitung:ZEIT Das kritische Denken des Autorenteams kommt ohne Feindbild nicht aus.

Abb. 8: Syntaktische Relation (auskommen_ohne, Feindbild) mit dem Relationstyp: V_PP

5.2 Zur Qualität des Wortprofils

Im vorangegangenen Abschnitt wurden die verschiedenen Nutzungsmöglichkeiten des Wortprofils erläutert. Noch nicht angeschnitten wurde die lexikographische Qualität des Wortprofils. Aufgrund des vom Korpus abgedeckten Zeitraums ist hierfür der Vergleich mit einem großen einsprachigen deutschen Gegenwartswörterbuch naheliegend. Für den Vergleich haben wir zwei große Wörterbücher herangezogen: das große Wörterbuch der deutschen Sprache in 10 Bänden des Dudenverlags ([GWDS]) und das Wörterbuch der deutschen Gegenwartssprache (WDG). Aus Platzgründen soll der Vergleich an dieser Stelle nur für ein Wort demonstriert werden, nämlich für das Adjektiv *grau*. Dieses Adjektiv haben wir ausgewählt, weil es häufig genug für ein ausgeprägtes Wortprofil ist, weil es mehrere Lesarten hat und weil es keine grundsätzlichen Bedeutungsveränderungen in den letzten Jahren erfahren hat. Eine ausführliche Darstellung aller Vergleichsparameter findet sich in Geyken (2011). Zusammengefasst hier noch einmal die wichtigsten Schlüsse, die sich aus dem Beispiel *grau* ableiten lassen. Wortprofile können im Vergleich zu großen einsprachigen Wörterbüchern ein Vielfaches der syntaktischen Relationen zu einem Wort enthalten. So verzeichnet das Wortprofil zu *grau* knapp 400 verschiedene Relationen, wohingegen das WDG 43, das GWDS nur 25 Wortverbindungen aufführen. Dies schlägt sich auch im direkten Vergleich nieder: Mit einer hohen statistischen Signifikanz (Salienz s>5) enthält das Wortprofil mehr als 20 lexikographisch relevante Beispiele, die nicht im WDG verzeichnet sind. Bei einer Salienz von unter fünf (s<5) im Wortprofil nimmt die Dichte der lexikographisch relevanten Relationen stark ab: von den knapp 200 syntaktischen Relationen (s<5) sind lediglich fünf als lexikographisch relevant einzustufen. Erstaunlich ist zunächst, dass das Wortprofil nur etwa 70 % der im Wörterbuch verzeichneten Wortverbindungen als syntaktische Relation enthält. Zu den meisten dieser fehlenden Verbindungen führt das Wortprofil jedoch gebräuchlichere Alternativen auf; in anderen Fällen existiert die Wortverbindung nur als Literaturzitat. Man findet im Wortprofil auch eine zusätzliche Lesart, die zwar nicht im WDG, jedoch im GWDS verzeichnet ist. Das Wortprofil (WP) hat aber hier die gebräuchlicheren Beispiele: *grauer Kapitalmarkt* oder *grauer Markt* (WP) statt *graue Händler* oder *graues Material* (GWDS). Schließlich, das zeigt das Beispiel *graue Theorie*, findet man mit dem Wortprofil zahlreiche authentische Kontexte und Verwendungen, die von dem einzigen dazu im WDG aufgeführten Goethezitat (*grau, mein Freund, ist alle Theorie*) abweichen. Eine Konstruktion übrigens, die in das GWDS nicht mehr aufgenommen wurde. In Abschnitt 6 wird gezeigt, dass sich diese positive Beurteilung des Wortprofils für die Probeartikel zum DWDS-Wörterbuch übertragen lässt.

5.3 DWDS-Wortprofil und lexikographische Intuition

Die Beschreibung der Kollokationen im DWDS-Wörterbuch basiert derzeit auf einer Mischung von aus dem Wortprofil extrahierten Relationen und Kompetenzkollokationen. Dass Kompetenzkollokationen auf absehbare Zeit das Wortprofil ergänzen müssen, ergibt sich aus der Tatsache, dass Kollokation und statistische Kookkurrenz zwar korrelieren, aber nicht identisch sein müssen. So kann es durchaus Kollokationen geben, die sich in den überwiegend schriftlichen Korpora des DWDS kaum niederschlagen und somit auch nicht im Wortprofil extrahiert werden können. Ein Beispiel für solch eine niedrigfrequente Kollokation liefert beispielsweise die Kollokationsbasis *Äußeres* mit den von Hausmann erwähnten Kollokatoren *angenehm*, *gepflegt*, *attraktiv*, *ansprechend* (alle im Wortprofil) und *einnehmend* (nicht im Wortprofil). Damit zusammenhängend ist die Tatsache, dass die im Wortprofil aufgelisteten Kollokationskandidaten Ergebnisse der hiermit dokumentierten Diskurse, weniger jedoch Ergebnis sprachlicher Möglichkeiten oder Präferenzen darstellen. Es gibt im Wortprofil bei-

spielsweise *konservative* Blogger, aber keine *progressiven*, weil der *konservative Blogger* der im Diskurs markierte Fall ist. Es gibt *erfolgreiche* Blogs, aber keine *langweiligen* oder *erfolglosen*, weil es sich über Letztere nicht zu reden lohnt.

6. Erprobung der Kollokationsanalyse

In einer Erprobungsphase wurden 136 Stichwörter lexikographisch für das DWDS-Wörterbuch ausgearbeitet. Die Ausarbeitung orientierte sich formal und inhaltlich an der strukturierten Version des elektronischen WDG, da die Probeeinträge zusammen mit der Grundsubstanz in das DWDS-Wörterbuch einfließen sollen. Kriterium für die Auswahl dieser Stichwörter war, dass sie im WDG nicht belegt sind, in heutigen Korpora jedoch hochfrequent sind und somit Kandidaten für Volleinträge im DWDS-Wörterbuch darstellen. Bei der Auswertung dieser Arbeit soll im Folgenden nur auf die Kollokationen Bezug genommen werden. Bei den ausgearbeiteten 136 Einträgen wurden 33„hausmannsche“ Kollokationen (Typ1) und 402 Kollokationen vom Typ2 in den Einträgen vermerkt. Nur ein relativ geringer Anteil, nämlich insgesamt 27 Kollokationen, konnte nicht aus dem Wortprofil extrahiert werden und wurde als Kompetenzkollokation vermerkt. Beispiele hierfür lassen sich aus allen Relationstypen finden:

- Adjektiv-Nomen (4 Kollokationen): eloquenter Schreibstil, inflatorische Verwendung, tägliche Charterflüge
- Objekt-Verb (14): Übertragungsrechte makeln, Gebäude observieren, Countdown abbrechen
- Verb-Adverb (4): dauerhaft archivieren, eloquent vertreten
- Präpositionalobjekt-Verb (1): zum Administrator ernennen
- Subjekt-Verb (3): der Jet startet, setzt auf; das Kraftwerk emittiert

7. Ausblick

Die Beschreibung der Kollokationen im DWDS-Wörterbuch basiert derzeit – und wohl auch noch auf absehbare Zeit – auf einer Mischung von aus dem Wortprofil extrahierten Relationen und Kompetenzkollokationen.

Eine weitere Verbesserung des Wortprofils ist auf drei Ebenen erreichbar: erstens auf der Ebene der Relationsextraktion mit Hilfe eines leistungsfähigeren Syntaxparsers. Diese Arbeiten, die insbesondere eine Verbesserung im Bereich der Verb-Relationen betreffen, sind bereits implementiert und Teil der nächsten Wortprofil-Version (DWDS-Wortprofil 2012). Auch durch die Variierung der statistischen Maße lassen sich Veränderungen und Verbesserungen der extrahierten Wortprofil-Relationen erzielen. So scheint die Ersetzung des bislang eingesetzten Salienzmaßes (Geyken et al. 2009) durch Dice-Koeffizienten (Rychly 2008) zu verbesserten Ergebnissen bei absolut gesehen hochfrequenten Wörtern zu führen, die mit dem Kollokator jedoch nicht häufiger als erwartbar auftauchen. Die dritte Ebene, auf der Verbesserungen der Qualität der extrahierten Wortprofile erzielt werden können, betrifft die Korpora. Die Korpora, die als Datengrundlage des DWDS-Wortprofils dienen, sind grundsätzlich frei wählbar. Die Zusammensetzung und Größe der Korpora spielen für das Wortprofil jedoch eine wichtige Rolle. Diese ist insofern relevant, als die extrahierten syntaktischen Relationen die im Korpus vorkommenden syntaktischen Nachbarn des Wortes widerspiegeln. Daher erhöht ein breit gestreutes, nach Textsorten ausgewogenes Korpus, ein sogenanntes allgemein-

sprachliches Referenzkorpus, die Qualität des Wortprofils in Bezug auf die allgemeinsprachliche Aussagekraft. Spezialkorpora oder spezielle Zeitungskorpora werden somit andere Wortprofile liefern als Referenzkorpora. Auch die Korpusgröße hat einen großen Einfluss auf die Wortprofile, denn Wortprofile sind in der Regel nur aussagekräftig, wenn das Lemma wenigstens 500, besser jedoch 1.000 Mal im Korpus auftaucht (siehe auch Ivanova et al. 2008). Unter dieser Zahl ist die Aussagekraft eines Wortprofils nur begrenzt, da viele syntaktische Relationen dann in der Regel nur ein oder zwei Mal vorkommen und somit kaum nachweisbar ist, dass es sich bei den extrahierten syntaktischen Relationen um typische Beispiele und nicht um Zufallsfunde handelt. Insofern spielt auch die absolute Korpusgröße eine Rolle, als sich mit wachsender Korpusgröße auch die Anzahl der verschiedenen Wörter erhöht, die hochfrequent im Korpus vorkommen. Dabei stellt sich heraus, dass eine Korpusgröße von 100 Millionen laufenden Textwörtern zu klein ist, um eine für die zu erwartende Benutzungssituation ausreichende Anzahl von Wortprofilen zu extrahieren. So gibt es beispielsweise im 100 Millionen Textwörter umfassenden DWDS-Kernkorpus nur etwa 5.000 Lemmata, die mehr als 1.000 Mal vorkommen. Bei dem 500 Millionen Textwörter großen Korpus, welches derzeit für das Wortprofil_2010 verwendet wird, gelangt man immerhin auf 15.000 Lemmata, die wenigstens 1.000 Mal im Korpus belegt sind. Wenn man die Schwellenwerte für die Mindestanzahl von Kontexten von vier auf drei senkt, verdoppelt sich in etwa die Anzahl der Lemmata. Dennoch ist auch diese Anzahl eine zu geringe Basis für ein umfangreiches einsprachiges Wörterbuch. Eine weitere Erhöhung der Korpusgrundlage um die weiteren Texte des DWDS-Korpus in einem Umfang von 2 Milliarden Textwörtern ist somit einer der nächsten geplanten Schritte.

8. Literatur

Didakowski, Jörg (2007): SynCoP – Combining syntactic tagging with chunking using WFSTs. In: Proceedings of FSMNLP 2007. Potsdam, S. 107-118.

Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century.In: Fellbaum, Christiane (Hg.): Collocations and Idioms: Linguistic, lexicographic, and computational aspects. London, S. 23-41.

Geyken, Alexander (2011): Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora. In:Abel, Andrea/Zanin, Renata (Hg.). Korpora in Lehre und Forschung. Bozen/Bolzano, S. 115-137.

Geyken, Alexander/Didakowski, Jörg/Siebert, Alexander (2009):Generation of word profiles for large German corpora. In: Kawaguchi, Yuji/Minegishi, Makoto/Durand, Jacques (Hg.): Corpus Analysis and Variation in Linguistics. Amsterdam, S.141-157.

[GWDS] Duden – Das große Wörterbuch der deutschen Sprache in 10 Bänden (1999). Mannheim. 3. Auflage.

Hausmann, Franz-Josef (1984): Wortschatzlernen ist Kollokationslernen. In: Praxis des neusprachlichen Unterrichts. 31. Jg. (1984), S. 395-406.

Herold, Axel (2011): Retrodigitalisierung und Modellierung des Wörterbuchs der deutschen Gegenwartssprache. In: Krafft, Andreas/Spiegel, Carmen (Hg.): Sprachliche Förderung und Weiterbildung transdisziplinär. Frankfurt/M. u.a., S. 197-213. (=Forum angewandte Linguistik 51)

Herold, Axel / Geyken, Alexander (2008): Adaptive word sense views for the dictionary database eWDG: The case of definition assignment. In: Storrer, Angelika/Geyken, Alexander/Siebert, Alexander/Würzner, Kay-Michael (Hg.): Text resources and lexical knowledge (TTCP 8). Berlin, S. 209-221.

Ivanova, Kremena/Heid, Ulrich/Schulte im Walde, Sabine/Kilgarriff, Adam/Pomikálek, Jan (2008): Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case. In: Proceedings of the 6th Conference on Language Resources and Evaluation. Marrakesch, Marokko, paper no. 537.

Klein, Wolfgang/Geyken, Alexander (2010): Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In: Heid, Ulrich/Schierholz, Stefan/Schweickard, Wolfgang/Wiegand, Herbert Ernst/Gouws, Rufus H./Wolski, Werner (Hg.):Lexikographica. Berlin/New York, S. 79-93.

Rychly, Pavel (2008): A Lexicographer-Friendly Association Score. In: Sojka, Petr /Horák, Aleš (Hg.): Proceedings of the 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2008. Brno, S. 6-9.

[WDG] Klappenbach, Ruth/Steinitz, Wolfgang (Hg.) (1964-1977): Wörterbuch der deutschen Gegenwartssprache. Berlin.

TEI-basierte Modellierung von Retrodigitalisaten (am Beispiel des Trierer Wörterbuchnetzes)

Vera Hildenbrandt hildenbr@uni-trier.de, Tel.: +49 651 201-3790

1. Einführung

Mit den Möglichkeiten des World Wide Web und den technologischen Entwicklungen seit Mitte der neunziger Jahre des 20. Jahrhunderts haben Stichwörter wie Performanz, Zugriffsmöglichkeiten, Interoperabilität, Hypertextualisierung, Langzeitarchivierung oder Nachhaltigkeit längst Einzug in die lexikographisch-lexikologische Diskussion gehalten und den lexikographischen Prozess, aber auch die Ansprüche von Nutzern an Wörterbücher verändert. Es stellen sich u.a. folgende Fragen: Wie können Wörterbücher im Internet präsentiert werden? Welche Zugriffsmöglichkeiten auf ihre Inhalte können geboten werden? Wie können Wörterbuchdaten sinnvoll miteinander vernetzt werden? Wie können sie nachhaltig vorgehalten und langfristig archiviert werden? Wie müssen Wörterbücher aufbereitet werden, um dauerhaft flexible vernetzte Präsentations- und Recherchemöglichkeiten zu schaffen?

Diese Fragen stellen den Bearbeiter von Wörterbüchern, die nicht von Anfang an digital publiziert, sondern erst durch Retrodigitalisierung ins elektronische Medium überführt werden, vor besondere Herausforderungen. Wie kann der Spagat zwischen Web-Technologien und Wörterbüchern, die ursprünglich nicht für die Internetpublikation konzipiert wurden, gelingen?Wie können retrodigitalisierte Wörterbücher so modelliert werden, dass die darin gebotenen lexikographischen Informationen auf verschiedene Weise präsentierbar und auf innovative Weise recherchierbar werden?

Fragen wie diese führten Ende 1987 unter der Ägide der Association for Computers and the Humanities, der Association for Computational Linguistics und der Association for Literary and Linguistic Computing zur Gründung der Text Encoding Initiative (TEI). Als Initiative der Wissenschaft für die Wissenschaft verfolgt die TEI seitdem das Ziel, hard- und software-unabhängige Methoden für die Codierung, den Austausch und die langfristige Archivierung geisteswissenschaftlicher Daten zu entwickeln. Im Jahr 1994 veröffentlichte die TEI die erste offizielle Version ihrer *Guidelines for the Encoding and Interchange of Machine-Readable Texts* (TEI P3),[9] die, basierend auf der Metasprache SGML, Codierungsregeln und -empfehlungen für eine Reihe verschiedener Textsorten und eine Vielzahl von Anwendungen vorschlug. Diese Regeln und Empfehlungen wurden seither ständig überarbeitet und erweitert; im November 2007 wurde die zurzeit aktuelle, XML-basierte und die Vorteile von XML-Schema nutzende Version P5 der Guidelines herausgegeben.[10] Die TEI-Guidelines beschreiben somit nicht nur ein in den Geisteswissenschaften mittlerweile sehr gängiges Dokumentenformat für Datencodierung und -austausch, sondern sie sind auch konform mit aktuellen Standards. Auf diese Weise können beispielsweise andere XML-Sprachen (z.B. MathML, SVG) ebenso in TEI-Dokumenten verwendet werden, wie TEI in anderen XML-Sprachen (z.B. MEI) genutzt werden kann.

[9] Die 1990 erschienene erste Version der Richtlinien (TEI P1) und die 1992 veröffentlichte Version P2 hatten nach dem Selbstverständnis der Verfasser keinen offiziellen, sondern „Entwurfscharakter“ (vgl. Schmidt 1997, S. 351).

[10] Vgl. zur Geschichte der Text Encoding Initiative die von der TEI selbst gebotene *History* unter http://www.tei-c.org/About/history.xml, zu den Anfängen außerdem Schmidt (1997), S. 351 f.

Die TEI-Guidelines wurden zur Modellierung der im Trierer Wörterbuchnetz versammelten retrospektiv digitalisierten Wörterbücher herangezogen. Im Folgenden soll eine allgemeine kurze Einführung in das TEI-Modul für Wörterbücher verdeutlichen, aus welchen Gründen das Codierungsschema der TEI in Trier eingesetzt wurde, bevor anschließend anhand einiger konkreter Beispiele die Vorzüge und Schwierigkeiten, die bei der Anwendung der TEI-Richtlinien auf retrodigitalisierte lexikographische Ressourcen auftreten, dargestellt werden.

2. Das TEI-Modul für Wörterbücher

An den Guidelines der TEI arbeiten seit der Gründung der Initiative Geistes- und Kulturwissenschaftler verschiedener Fachdisziplinen. Rückmeldungen der Anwender und die Arbeit verschiedener Special-Interest-Groups tragen zur ständigen praxisnahen Weiterentwicklung eines vielseitigen Regelwerks bei, das Standards setzt, den Nutzern aber dennoch Freiräume lässt.[11] Das Codierungsschema der TEI besteht aus verschiedenen Modulen, die jeweils eine bestimmte Anzahl von XML-Elementen (Tags) deklarieren, die gegebenenfalls durch Attribute spezifiziert werden.[12] Neben einem sogenannten „Core-Tagset", das in jedem TEI-Dokument zur Verfügung steht, gibt es einige „Base-Tagsets", so unter anderem ein Tagset zur Auszeichnung von Prosatexten, eines zur Auszeichnung von Verstexten, eines zur Auszeichnung von Manuskripten und eben auch eines zur Auszeichnung von Wörterbüchern.

Die Entwicklung der Richtlinien für die Codierung von Wörterbüchern stellte die zuständige Arbeitsgruppe vor besondere Herausforderungen, denn es galt, ein Schema zu entwickeln, das so allgemein ist, dass es für eine Vielzahl von Wörterbüchern anwendbar ist und zugleich die Besonderheiten des einzelnen Wörterbuchs abdeckt.[13] Das Schema muss berücksichtigen, dass Wörterbücher – auch typographisch – zu den komplexesten Textsorten gehören, die von der TEI behandelt werden, dass sie sehr stark und tief strukturiert und die enthaltenen Informationen sehr verdichtet sein können und dass ebenso die Struktur von Artikeln verschiedener Wörterbücher wie auch die Struktur von Artikeln innerhalb eines Wörterbuchs variieren kann. Etymologische Informationen etwa können an unterschiedlichen Stellen und auf unterschiedlichen Hierarchieebenen innerhalb des Wörterbuchartikels erscheinen, oder Informationen auf einer niedrigeren Hierarchieebene können Informationen auf einer höheren Hierarchieebene ersetzen. Das heißt, dass sich dieselben Informationsklassen auf allen Hierarchieebenen eines Wörterbuchartikels finden können. Die TEI-Richtlinien für Wörterbücher müssen also zulassen, dass alle XML-Elemente auf allen Ebenen eines Wörterbuchartikels vorkommen können und dass im Umkehrschluss alle Elemente, die eine Artikelebene kennzeichnen, denselben Inhalt haben dürfen.[14]

Aus der Bewältigung dieser Herausforderungen durch die Entwickler des Wörterbuchmoduls der TEI, das heißt aus der Umsetzung der Erkenntnisse über die mögliche(n) Struktur(en) von Wörterbüchern und ihren Artikeln in ein Regelwerk ergibt sich ein Teil der Gründe, die für den Einsatz der TEI-Guidelines bei der Modellierung von Retrodigitalisaten sprechen: Erstens

[11] Vgl. dazu Jannidis (1997), S. 153/154.

[12] Auf eine Beschreibung des Aufbaus einer TEI-Datei sei an dieser Stelle mit Verweis auf Jannidis (1997), S. 155-158, verzichtet. Obwohl bereits 1997 erschienen, darf die allgemeine, knappe Einführung in die Teile einer TEI-Datei auch heute noch Gültigkeit beanspruchen und ist gerade für Neuanwender sehr gut verständlich.

[13] Entwickelt wurde das Wörterbuchtagset zunächst für Wörterbücher mittlerer Größe und westlicher Sprachen wie den *Petit Larousse*, den *Petit Robert* und das *Collins English Dictionary* (vgl. Ide/Véronis 1996b, S. 170).

[14] Zu den Herausforderungen, die bei der Entwicklung von Richtlinien für die Auszeichnung von Wörterbuchartikeln zu bewältigen waren, vgl. Ide/Véronis (1996a) und Ide/Véronis (1996b), S. 173-175.

sind die Richtlinien ganz bewusst so allgemein gehalten, dass sie auf unterschiedlich konzipierte Wörterbücher angewendet werden können. Zweitens ermöglichen sie benutzerdefinierte Erweiterungen. Drittens erlauben sie eine nur geringe Explizitheit der Auszeichnung, das heißt, nicht jeder Wörterbuchartikel muss bis in seine feinsten Verästelungen hinein codiert sein, bevor eine elektronische Publikation vorgenommen werden kann. Viertens sind durch das Befolgen der mittlerweile international anerkannten und bei der Codierung einer Vielzahl von historischen Wörterbüchern eingesetzten Guidelines gute Voraussetzungen für die Interoperabilität der Wörterbuchdaten und damit für die Vernetzung mit anderen Nachschlagewerken geschaffen. Fünftens kann mittels der Richtlinien nicht nur die Oberflächenstruktur der Wörterbücher (Layout, Typographie, lineare Abfolge von Zeichen und Informationen) codiert werden, wodurch es beispielsweise möglich wird, in der Präsentation einer vor der digitalen Version existierenden gedruckten Fassung, wie sie bei Retrodigitalisaten immer gegeben ist, treu zu bleiben. Vielmehr kann das Wörterbuch durch die XML-basierte Codierung auch als Datenbank betrachtet werden, die der Nachschlagende nicht von A bis Z liest, sondern in der er gezielt sucht und selektiert.[15] Die TEI-Codierung von Wörterbüchern schafft folglich nicht nur die nötigen Voraussetzungen für eine flexible Präsentation der Wörterbuchinhalte, sondern bereitet auch verschiedene Zugriffsmöglichkeiten auf diese Inhalte vor. Und sechstens schließlich wurden die Guidelines bereits in vielen Projekten erfolgreich erprobt und nicht zuletzt bei der Codierung (retrodigitalisierter) älterer Wörterbücher eingesetzt.[16]

Das TEI-Modul „Dictionaries“ erlaubt die Abbildung der Makro- und Mikrostruktur von lexikographischen Ressourcen verschiedener Art. Für die Makrostruktur von gedruckten Wörterbüchern sind analog zu anderen gedruckten Texten die Tags <text>, <front>, <body>, <back> und vorgesehen. Der Wörterbuchartikel selbst wird in der Regel durch das Element <entry> bezeichnet; Artikel, die in höchstem Maße unstrukturiert sind und die mit dem <entry>-Tag verbundenen Bedingungen nicht erfüllen, dürfen in Ausnahmefällen durch den <entryFree>-Tag gekennzeichnet werden. Neben diesen Elementen sind weitere die Hierarchie eines Wörterbuchartikels abbildende Tags der Homographentag <hom>, der zur Codierung aller Informationen, die sich auf einen Homographen innerhalb eines Wörterbuchartikels beziehen, herangezogen wird, und das Element <sense>, das zur Markierung aller auf eine bestimmte Bedeutung des Stichwortes bezogenen Angaben eingesetzt wird. In diese größeren hierarchisierenden Struktureinheiten können dann weitere lexikographische Informationseinheiten eingeschachtelt sein. Hier sehen die TEI-Guidelines u.a. Tags vor zur Auszeichnung von Informationen über die geschriebene und gesprochene Form eines Stichwortes (<form>), von morphosyntaktischen Informationen (Wortart, Genus, Numerus, Kasus etc.) (<gramGrp>), von Bedeutungsangaben (<def>), von Zitaten inklusive bibliographischer Angaben (<cit>), von Angaben zum Gebrauch des Stichwortes (<usg>), von Querverweisen innerhalb des Wörterbuchs oder auf andere Werke (<xr>), von etymologischen Informationen (<etym>), von Verweisen auf Komposita oder Derivate des Stichworts (<re>) und schließlich zur Auszeichnung von Anmerkungen (<note>). Die verschiedenen durch diese Elemente jeweils gekennzeichneten Informationen, die die TEI als Hauptkonstituenten eines Wörterbuchartikels betrachtet, können durch weitere Tags feiner ausgezeichnet werden. So kann etwa bei den Informationen zu einem Stichwort weiter differenziert werden zwischen der orthographischen Form des Lemmas (<orth>), Angaben zur Aussprache (<pron>), zur Silbentrennung (<syll>) usw. Zusätzlich können weitere, in allen TEI-Dokumenten verfügbare Elemente zur Kennzeichnung von typographischen und Layout-Merkmalen – zum Beispiel des Zeilen-

[15] Vgl. Ide/Véronis (1996b), S. 174.

[16] Vgl. etwa den Bericht von Susan Rennie über die Codierung des *Scottish National Dictionary* (Rennie 2000) sowie die Liste der Projekte, die das Codierungsschema der TEI nutzen, unter http://www.tei-c.org/Activities/Projects/.

umbruchs (<lb>), des Seitenumbruchs (<pb>) oder des Spaltenwechsels (<cb>) – genutzt werden, um den Bezug zur Printfassung zu erhalten. Doch sollen die allgemeinen Ausführungen über die Richtlinien der TEI für die Codierung von Wörterbüchern an dieser Stelle nicht weiter vertieft werden,[17] vielmehr soll im Folgenden die Anwendung der Guidelines bei der Auszeichnung retrodigitalisierter Wörterbücher am Beispiel des Trierer Wörterbuchnetzes im Zentrum stehen.

3. Codierung retrodigitalisierter Wörterbücher nach den Richtlinien der TEI

Im Trierer Wörterbuchnetz ist zwischenzeitlich eine Reihe verschiedener Nachschlagewerke versammelt, von denen die Mehrzahl erst lange nach Erscheinen der gedruckten Version ins digitale Medium überführt wurde. Mit Förderung durch die Deutsche Forschungsgemeinschaft (DFG) wurden am Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier (Trier Center for Digital Humanities) seit 1997 das *Mittelhochdeutsche Wörterbuch* von Georg Friedrich Benecke, Wilhelm Müller und Friedrich Zarncke (BMZ, 1854-1866), das *Mittelhochdeutsche Handwörterbuch* von Matthias Lexer (Lexer, 1872-1878), Lexers Nachträge zum *Handwörterbuch*, das *Findebuch zum mittelhochdeutschen Wortschatz* (Findebuch, erarbeitet zwischen 1986 und 1992), die Erstbearbeitung des *Deutschen Wörterbuchs* von Jacob und Wilhelm Grimm ([1]DWb, 1852-1960), das *Pfälzische Wörterbuch* (PfälzWb, 1965-1997), das *Rheinische Wörterbuch* (RheinWb, 1928 und 1971), das *Wörterbuch der deutsch-lothringischen Mundarten* (LothrWb, 1909) und das *Wörterbuch der elsässischen Mundarten* (ElsWb, 1899-1907) digitalisiert. Ebenfalls mit Förderung durch die DFG wurden parallel zur Ausarbeitung der weiteren Bände des Wörterbuchs die zwischen 1966 und 1998 veröffentlichten ersten drei Bände des *Goethe-Wörterbuchs* (GWb) digitalisiert und im Internet publiziert. Zu diesen Ressourcen kommen mit *Meyers Großem Konversationslexikon* und Adelungs *Grammatisch-Kritischem Wörterbuch der Hochdeutschen Mundart* weitere Nachschlagewerke, deren Daten im Rahmen des vom Bundesministerium für Bildung und Forschung geförderten Verbundprojektes TextGrid erworben werden konnten, sowie eine Reihe externer Ressourcen (*Mittelhochdeutsches Wörterbuch*, Mittelhochdeutsche Begriffsdatenbank, *Deutsches Rechtswörterbuch*, *Oekonomische Encyklopädie* von Krünitz, *Lexikon der Luxemburger Umgangssprache*, *Wörterbuch der luxemburgischen Mundart*, *Luxemburger Wörterbuch*), auf die aus dem Trierer Wörterbuchnetz und vice versa verlinkt wird. Im Rahmen dieses Beitrags stehen vor allem die Wörterbücher im Fokus, deren digitale Version im Rahmen von DFG-geförderten Projekten entstanden ist – also BMZ, Lexer, Findebuch, [1]DWb, PfälzWb, RheinWb, LothrWb, ElsWb und GWb.

Gemeinsam ist diesen Wörterbüchern, dass sie – es wurde oben bereits erwähnt und wird zudem durch die Erscheinungsdaten deutlich – ausschließlich im Druck veröffentlicht und mit Ausnahme des GWb bereits abgeschlossen wurden, als noch nicht an eine elektronische Version und deren Möglichkeiten gedacht wurde. Teilweise sind diese Wörterbücher über einen sehr langen Zeitraum hinweg entstanden; als prominentestes Beispiel muss hier sicherlich das [1]DWb angeführt werden, dessen erste Lieferung 1852 und dessen letzte Lieferung im Jahr 1960 erschien. Der in diesen Zeiträumen erfolgte Wandel sprachwissenschaftlicher und lexikographischer Grundsätze, aber auch den Lexikographen auferlegte Straffungskonzepte, die einen rascheren Abschluss der Unternehmen bewirken sollten, führten im Verlauf der Erarbeitung der Nachschlagewerke zu Änderungen in der lexikographischen Praxis. Auch hier ragt

[17] Für weiterführende Informationen sei auf Kapitel 9 „Dictionaries“ der TEI-Guidelines verwiesen. Eine PDF-Datei ist zugänglich unter: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

das [1]DWb aus der Riege der genannten Wörterbücher hervor, denn bereits „[v]on Beginn an, also seit Jacob und Wilhelm Grimm, gibt es ungewollte und gewollte Differenzen im Aufbau und in der Aussageleistung der Artikel." (Schmidt 2004, S. 25) Auch eine erste Reorganisation der Wörterbucharbeit im Jahr 1908 sowie eine zweite im Jahr 1930 änderten nur wenig an der Heterogenität in der Artikelstruktur.[18] Aber auch ein jüngeres Wörterbuch wie das GWb wurde Straffungskonzepten unterworfen, die zu Modifikationen in der lexikographischen Praxis führten.[19] Gemeinsam ist den hier genauer betrachteten Nachschlagewerken also auch, dass es sich um in mehr oder minder geringem Maße standardisierte Wörterbücher handelt, deren Artikel sich – etwa im Fall von BMZ und [1]DWb – durch einen eher diskursiven Stil und zum Teil durch eine wenig stringente Artikelstruktur auszeichnen. Durch die Diskursivität in der Darbietung der Informationen wird es schwierig bis unmöglich, die Zugehörigkeit der Informationen zu einer bestimmten Informationsklasse programmgestützt zu bestimmen. Zudem werden Informationen nicht immer in derselben Reihenfolge angeboten. Im BMZ beispielsweise können Angaben zur Morphologie, zur Bedeutung und zur Etymologie in beliebiger Abfolge und an beliebigen Stellen eines Wörterbuchartikels vorkommen, unter Umständen können sie auch mehrmals innerhalb desselben Artikels auftreten.[20] Und gemeinsam sind den Wörterbüchern schließlich auch gewisse Inkonsistenzen – so zum Beispiel bei den zitierten Siglen oder den Einleitungen eines Querverweises.

Diese Wörterbücher mit ihrer jeweils spezifischen Geschichte, ihren Gemeinsamkeiten und ihren Problemen wurden retrodigitalisiert mit dem Ziel, sie als Open-Access-Publikationen im Internet anzubieten. Dabei sollten sie zum einen als elektronische Abbilder der gedruckten Versionen ins Internet gestellt werden, um auch aus der digitalen Publikation heraus die Printausgaben referenzier- und zitierbar zu halten. Das heißt, dass bei der Aufbereitung und Auszeichnung der Daten weder der Wörterbuchtext an sich noch die Struktur der Artikel verändert werden durfte. Zum anderen sollten die Wörterbücher jedoch als Datenbank begriffen werden, in der bestimmte lexikographische Informationseinheiten gezielt such- und durchsuchbar sein sollten. Darüber hinaus sollten zumindest die bereits im Druck eng aufeinander bezogenen Wörterbücher, also zum einen BMZ, Lexer und Findebuch und zum anderen PfälzWb, RheinWb, LothrWb und ElsWb, in der digitalen Version miteinander vernetzt werden. Zur Umsetzung dieser Ziele standen jeweils nur ein begrenzter Zeitraum und ein begrenztes Budget zur Verfügung. Dies bedeutete, dass nach der Retrodigitalisierung der Wörterbücher, die im Double-Keying-Verfahren erfolgte, der Prozess der Auszeichnung auf möglichst ökonomische Weise, das heißt weitgehend automatisiert durchgeführt werden musste. Es wurden daher Programmroutinen entwickelt, die – auf den bei der Eingabe miterfassten typographischen Merkmalen aufsetzend – das TEI-Markup nach der Top-down-Methode in die Wörterbuchdaten einbrachten. Das heißt, zuerst wurden die oberen Ebenen der Artikelstruktur ausgezeichnet und von hier aus dann stufenweise die weiteren lexikographischen Informationseinheiten in Angriff genommen. Dabei wurde die Tatsache, dass die meisten der genannten Wörterbücher wenig standardisiert sind, zu einer Herausforderung für die computergestützte Auszeichnung.

[18] Eine kurzgefasste Geschichte des [1]DWb bieten Christmann/Hildenbrandt/Schares (2001), S. 14-20.

[19] So wurde etwa ab Band 3 des GWb die in den Verweisreihen der ersten beiden Bände gemachte Unterscheidung zwischen Synonymen und im weiteren Sinne bedeutungsverwandten Wörtern und Ausdrücken aufgegeben, und ab Lieferung IV,11 wurden die Artikelstruktur gestrafft sowie die Behandlung hochfrequenter Wörter (z.B. Präpositionen, Konjunktionen, Modalverben) radikal auf die Darstellung Goethescher Besonderheiten reduziert (vgl. Schmidt/Reinitzer/Kühlmann 2004).

[20] Vgl. zum Problem der geringen Standardisierung der BMZ-Artikel Burch/Fournier (2001), S. 140 ff.

Wie bei den Ausführungen über das Modul „Dictionaries“ der TEI-Guidelines dargestellt, wurde bei der Entwicklung eines Codierungsstandards für elektronische Wörterbücher zwar grundsätzlich berücksichtigt, dass Wörterbuchartikel in ihrer Struktur variieren können – so dürfen etwa die Elemente <form>, <gramGrp>, <sense> und <etym> ohne zwingend vorgeschriebene Reihenfolge wiederholt innerhalb eines <entry>-Tags vorkommen. Die Tatsache aber, dass bestimmte typographische Eigenschaften unterschiedliche Arten von Textelementen abbilden, das Fehlen eindeutiger struktureller Marker, Variantenreichtum in Bezug auf eine bestimmte Information und vor allem im Wörterbuch nur implizit gegebene Informationen, die sich von einem menschlichen Benutzer ohne Probleme erschließen lassen, stellen – einige Beispiele werden dies in der Folge verdeutlichen – eine große Hürde für ein maschinelles Markup dar. Für die in Trier digitalisierten Wörterbücher bedeutete dies, dass mittels computergestützter Verfahren innerhalb des jeweils zur Verfügung stehenden Bearbeitungszeitraums je nach Umfang und Struktur des Wörterbuchs unterschiedliche Auszeichnungstiefen erreicht werden konnten. Das [1]DWb mit seinen 16 Bänden in 32 Teilbänden, das auf nahezu 70.000 Spalten rund 320.000 Stichwörter in heterogen strukturierten und vielfach sehr diskursiven Wörterbuchartikeln behandelt, steht hier am einen Ende der Skala, das GWb, das in voraussichtlich neun Bänden rund 90.000 Stichwörter in relativ konsistent strukturierten Artikeln beschreiben wird, am anderen Ende. Anhand dieser beiden Wörterbücher seien daher im Folgenden die Möglichkeiten und Grenzen der computergestützten TEI-basierten Modellierung von Retrodigitalisaten demonstriert.

Buch **I** *das zu einer Werkeinheit gebundene od geheftete, geschriebene od gedruckte Literaturwerk* **1** *konkret, bes im Hinblick auf Format u Bindeart, Typographie u Ausstattung* **a** *als handschriftl Buch* Das schöngeschriebne B. 6,38 Vs 4 DivHafis B25,184,8 Lorsbach 31.1.15 B21,14,21 Christiane 28.7.09 *uö* **b** *als gedrucktes Buch; mehrf '(ein)gebundenes, ungebundenes, geheftetes, gefalztes, rohes B.', vereinzelt 'durchschossenes B.'* Wenn typographisch allgemach die Bücher sich steigern, darf wohl auch der Buchbinder ehrenvoll als Künstler hervortreten 49²,135,2 KLehmanns Buchbinderarb Zu der andern Ausgabe [*von HermDor*] bin ich ganz wohl mit der hierbey zurückkommenden lateinischen Schrift zufrieden, nur wünsche ich einen breiten Steg und überhaupt viel Rand, als die wahre Zierde jedes B-es B12,135,11 Böttiger 3.6.97 Der Druck [*von KuA*] nimmt sich wie bisher mit den wenigen Abänderungen sehr gut aus; alles kömmt auf's Papier an, woran denn freylich unsre meisten Hefte und Bücher kranken B41,248,12 FJFrommann 9.12.26 K die Kupferstecher tractiren alles was zu einem B-e gehört so leicht und lose B12,368,13 Schiller 2.12.97 A(Lerche 70) ChAVulpius 11.6.00 [G/Voigt BiblKomm] B9,38,25 CarlAug 8.10.88 30,87,25 ItR B10,335,16 Schiller 21.11.95 B8,277,16 Göschen [27.10.87] *uö* **c** *in sekundärer Verwendung als Aufbewahrungsort für empfindl Papiere, insbes zum Pressen u Bewahren von Pflanzen u Insekten* B25,338,13 Meyer 17.5.15 N6,336,9 Bryophyllum calycinum N6,419,3 MetamIns *uö* **2** *inhaltl*

Abb. 1: Der Anfang des Artikels *Buch* im GWb (Band 2, Spalte 923 ff.)

Der Beginn des Artikels *Buch* in der Printversion des GWb (vgl. Abb. 1) verdeutlicht, wie sehr sich die inhaltliche Strukturierung in der Typographie dieses Wörterbuchs widerspiegelt: Das Lemma ist halbfett gesetzt, die lexikographischen Interpretamente sind kursiv und halbfett gedruckt, die Belegzitate werden durch eine andere Schrifttype gekennzeichnet, die Belegstellenangaben durch einen kleineren Schriftgrad und eine geringere Zeichenbreite. Dank dieser sehr stark differenzierten typographischen Gestaltung und dem trotz der bereits mehr als 60 Jahre währenden Bearbeitungszeit sehr regelhaften Aufbau der Wörterbuchartikel konnte die TEI-konforme Codierung des GWb weitgehend automatisiert und vergleichsweise zügig durchgeführt werden, und es konnte eine beträchtliche Auszeichnungstiefe erreicht werden.

Dies wird illustriert durch die Präsentation des Wörterbuchs im Internet (vgl. Abb. 2). Die typographischen Merkmale der Druckfassung sind weitgehend erhalten, allerdings mit dem Unterschied, dass verschiedene Informationseinheiten in verschiedenen Farben markiert sind. Lexikographische Interpretamente sind rot, Belegzitate grün und Belegstellenangaben blau gehalten. Eine Präsentationsform, die Gloning und Schlaps bei ihren Vorüberlegungen zu möglichen *Prototypen für ein elektronisches Goethe-Wörterbuch* bedachten, ist damit umgesetzt: Es liegt „derselbe Datenbestand vor wie im gedruckten Wörterbuch, die Ansicht ist jedoch deutlich stärker strukturiert." (Gloning/Schlaps 1999, S. 26).

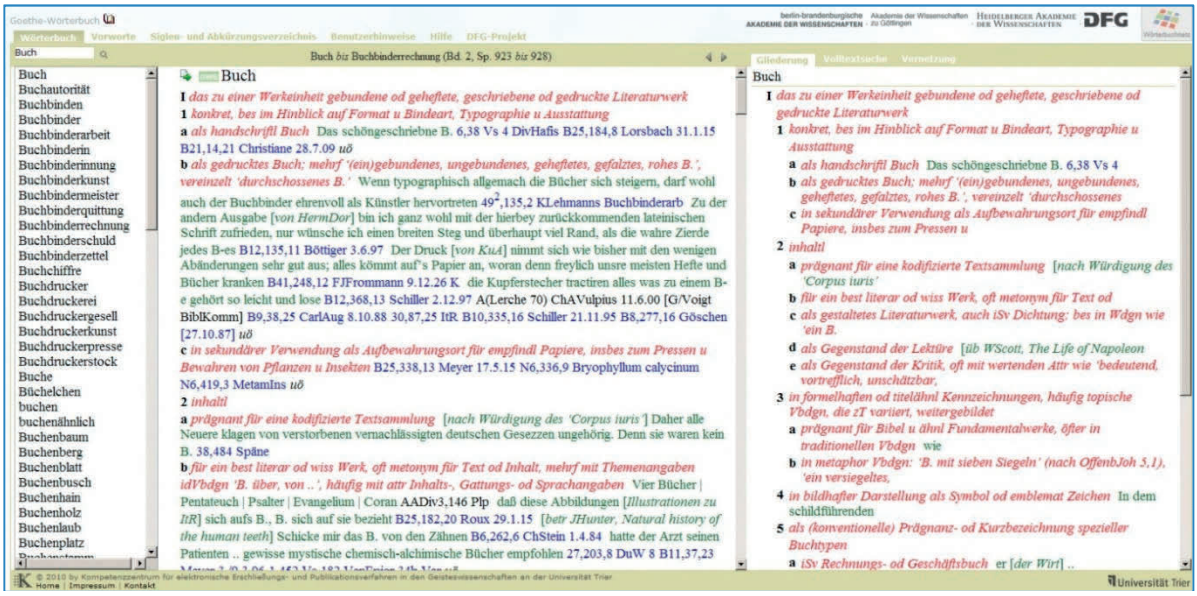


Abb. 2: Graphische Oberfläche des GWb im Internet

Zusätzlich werden im rechten Anzeigebereich die Wörterbuchartikel auf ihr hierarchisches Bedeutungsgerüst, also Lemma, Nummerierung und lexikographische Interpretamente, reduziert, im linken Anzeigebereich werden lediglich die Lemmata angeführt. Auf diese Weise vermag sich der Nutzer einen rascheren Überblick über die Artikelinhalte zu verschaffen. Neben diesen verschiedenen Präsentationsmöglichkeiten wurden aber auch verschiedene Recherchemöglichkeiten geschaffen (vgl. Abb. 3). Der Benutzer kann im Volltext suchen, er kann die Suche aber auch auf die Stichwörter, die Bedeutungserklärungen, die Belegzitate und die Belegstellenangaben beschränken, wobei die Kombination der verschiedenen Suchfelder ebenso erlaubt ist wie der Einsatz von Wildcards.

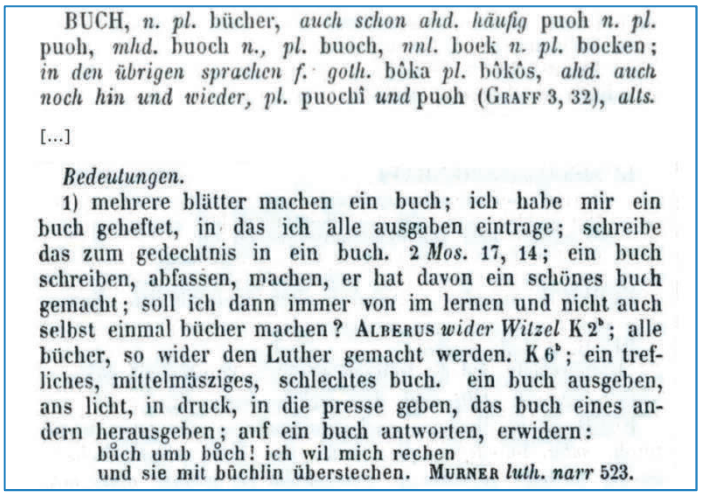
Abb. 3: Graphische Oberfläche des GWb im Internet mit Suchmaske

Diese verschiedenen Präsentations- und Recherchemöglichkeiten setzen voraus, „dass die entsprechenden Informationen [...] in expliziter Weise im Datenbestand enthalten sind.“ (Gloning/Welter 2001, S. 118). Gloning und Welter bringen diese Notwendigkeit auf den Punkt in dem knappen Satz: „What you mark is what you get“ (Gloning/Welter 2001, S. 128). Im GWb sind folgende Informationseinheiten markiert und damit gezielt „greifbar“: der Artikelkopf mit Lemma und Vorbemerkung, der Artikelkörper mit den verschiedenen Bedeutungsblöcken, Gliederungsmarken und lexikographischen Interpretamenten, Belegzitaten und Belegstellenangaben und der Artikelanhang mit dem Verweisblock, der auf Synonyme, Derivate und Komposita des Stichworts verweist, aber auch Anmerkungen enthalten kann (vgl. Abb. 4).

```
<entry xml:id="JB04816" n="20923.48">
      <form type="artkopf">
            <ref type="DWB" target="B2S0466Z79"></ref>
            <form type="lemma">Buch</form>
      </form>
      <sense rend="artkoerper">
            <sense n="I" level="2" value="20923.48">
                  <def rend="leitbem">das zu einer Werkeinheit gebundene od
                  geheftete, geschriebene od gedruckte Literaturwerk</def>
                  <sense n="1" level="3" value="20923.49">
                        <def rend="leitbem">konkret, bes im Hinblick auf Format u
                        Bindeart, Typographie u Ausstattung</def>
                        <sense n="a" level="4" value="20923.50">
                              <def rend="leitbem">als handschriftl Buch</def>
                              <cit>
                                    <quote>Das sch&ouml;ngeschriebne
                                    B.</quote>
                                    <bibl>6,38 Vs 4 DivHafis</bibl>
                              </cit>
[...]
      <sense rend="artanhang">
            <sense rend="verweisblock">
                  <xr type="derivblock">&derivsign;
                        <ref type="deriv" target="B1S0017Z54">ABC-Buch</ref>
[...]
                  <xr type="anmblock">
                        <note type="footnote" n="1">1) <note><hi rend="italic">Briefw
                        von Riemer,
[...]
                  <ref type="artikelautor" rend="H. U." n="Horst Umbach"></ref>
      </sense>
</entry>
```

Abb. 4: Auszug aus dem XML-codierten Artikel *Buch* des GWb

Eine so feinkörnige Modellierung der Daten konnte für das [1]DWb nicht erreicht werden, was unter anderem damit zusammenhängt, dass das Ausgangsmaterial, die gedruckte Fassung, weniger konsistent strukturiert und typographisch weit weniger differenziert ist als die des GWb. Im [1]DWb wechseln sich im Wesentlichen recte und kursiv gesetzte Passagen ab (vgl. Abb. 5). Lediglich die Lemmata, die Autornamen und die Verszitate heben sich durch Versalien, Kapitälchen bzw. Einrückung vom umgebenden Wörterbuchtext ab.



BUCH, *n. pl.* bücher, *auch schon ahd. häufig* puoh *n. pl.* puoh, *mhd.* buoch *n., pl.* buoch, *nnl.* boek *n. pl.* boeken; *in den übrigen sprachen f. goth.* bôka *pl.* bôkôs, *ahd. auch noch hin und wieder, pl.* puochî *und* puoh (GRAFF 3, 32), *alts.*

[...]

Bedeutungen.

1) mehrere blätter machen ein buch; ich habe mir ein buch geheftet, in das ich alle ausgaben eintrage; schreibe das zum gedechtnis in ein buch. 2 *Mos.* 17, 14; ein buch schreiben, abfassen, machen, er hat davon ein schönes buch gemacht; soll ich dann immer von im lernen und nicht auch selbst einmal bücher machen? ALBERUS *wider Witzel* K 2^{b}; alle bücher, so wider den Luther gemacht werden. K 6^{b}; ein treffliches, mittelmäsziges, schlechtes buch. ein buch ausgeben, ans licht, in druck, in die presse geben, das buch eines andern herausgeben; auf ein buch antworten, erwidern:

bůch umb bůch! ich wil mich rechen
und sie mit bůchlin überstechen. MURNER *luth. narr* 523.

Abb. 5: Auszug aus dem Artikel *Buch* im [1]DWb (Band 2, Spalte 466 ff.)

Maschinell lassen sich hier daher auch nur die Lemmata, die den Lemmata folgenden Wortartangaben, die Autornamen, die Verszitate und die Gliederungsmarken markieren (vgl. Abb. 6).

```
<entry xml:id="GB12398" n="2.0466.79">
    <form>
        <form type="lemma">buch</form>,
    </form>
    <gramGrp>
        <gram type="n">n.</gram>
    </gramGrp>
    <sense>
        <sense> <hi rend="italic">pl.</hi> b&uuml;cher, <hi rend="italic">auch schon
        ahd. h&auml;ufig</hi> puoh <hi rend="italic">n. pl.</hi> puoh,
        <hi rend="italic">mhd.</hi> buoch <hi rend="italic">n., pl.</hi> buoch,
        <hi rend="italic">nnl.</hi> boek <hi rend="italic">n. pl.</hi> boeken;
[...]
    </sense>
    <sense n="" level="1" value="2.0467.65"><title type="head">Bedeutungen.</title>
        <sense n="1)" level="2" value="2.0467.66"> mehrere bl&auml;tter machen ein
        buch; ich habe mir ein buch geheftet, in das ich alle ausgaben eintrage; schreibe
        das zum gedechtnis in ein buch.
        <title type="sigle">
            <bibl>
                <title type="bibel">
                    <title>2 Mos.</title>
                    <ref>17, 14</ref>
                </title>
            </bibl>
        </title>;
[...]
        <cit type="verse">
            <quote>b&oaboveu;ch umb b&oaboveu;ch! ich wil mich rechen<lb/>
            und sie mit b&uuml;chlin &uuml;berstechen.</quote>
            <bibl>
                <author rend="verse">Murner</author>
                <title>luth. narr</title> <ref>523</ref>.
            </bibl>
        </cit>
[...]
        </sense>
        <sense n="2)" level="2" value="2.0468.10"> b&uuml;cher lesen, einsehen,
[...]
        </sense>
    </sense>
</entry>
```

Abb. 6: Auszug aus dem XML-codierten Artikel *Buch* des [1]DWb

Weniger explizit markierte Informationen aber bedeuten eine im Vergleich zum GWb weniger strukturierte Artikeldarstellung (vgl. Abb. 7) und auch weniger Recherchemöglichkeiten.

Abb.7: Synoptische Darstellung des Artikels *Buch* im GWb und im [1]DWb

In der Online-Version des [1]DWb ist neben einer Volltextsuche lediglich eine auf die Stichwörter reduzierte Suche möglich (vgl. Abb. 8). Dennoch bieten bereits diese Suchmöglichkeiten in der elektronischen Version gegenüber dem Nachschlagen in der Buchausgabe Vorteile. Der Nutzer kann beispielsweise „auf viel leichtere Weise Wörter suchen, die zwar keinen eigenen Stichwortansatz haben, aber in den Belegen anderer Wörter enthalten sind [...].“ (Schmidt 2004, S. 27)[21].

Gleichwohl beeinflusst und erschwert die Konzeption des [1]DWb ebenso wie die anderer Wörterbücher die Modellierung retrodigitalisierter Daten. Dies sei im Folgenden anhand einiger weiterer Beispiele verdeutlicht.

[21] So stellte vor einigen Jahren ein Nutzer folgende Anfrage an das Retrodigitalisierungsprojekt: „In der 1821 von Gemeiner herausgegebenen ‚Regensburgischen Chronik‘ findet sich in einem Zitat der Begriff ‚thaem‘: ‚[...] es sey by der Marter des vor acht Jahren gemordeten Kindes ein solches Geschrey und Thaem gewesen [...]‘. Leider finde ich diesen Begriff im Grimmschen Wörterbuch nicht. Können Sie mir weiterhelfen?“ In der Druckausgabe findet der Nachschlagende ein in dieser Schreibung angesetztes Stichwort und damit dessen Bedeutung tatsächlich nicht ohne Weiteres. Über die Volltextsuche im Wörterbuchtext wird er im digitalen [1]DWb dagegen schnell fündig. Er stößt auf einen Beleg im Artikel DÄM GEDÄM und findet hier die Erklärung „*anschlagen der waffen, waffenlärm im kampf*“.

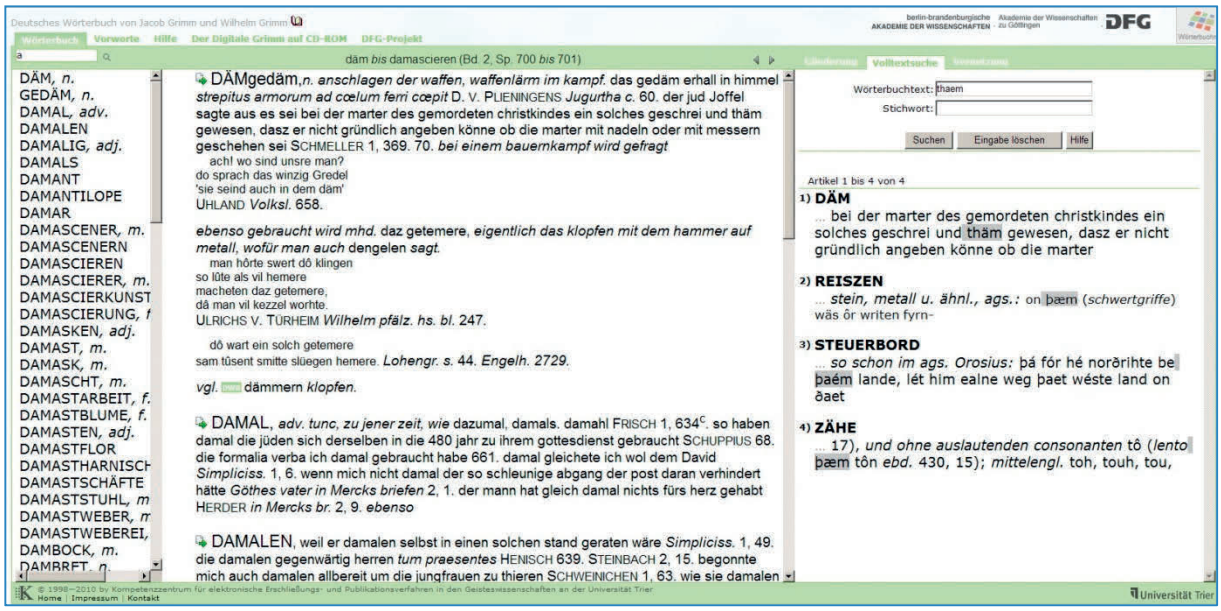


Abb. 8: Graphische Oberfläche des [1]DWb im Internet mit Suchmaske

Die Möglichkeit, die Artikelansicht auf das Gliederungsgerüst zu reduzieren, konnte im digitalen [1]DWb nur mit sehr hohem Aufwand realisiert werden, da das Gliederungsschema im [1]DWb im Vergleich beispielweise zu dem des GWb sehr viel weniger konsistent ist. Zwar können die Gliederungsmarken im [1]DWb automatisch ausfindig gemacht werden, da sie in der Regel etwas eingerückt am Beginn einer neuen Zeile stehen, doch ist es sehr schwierig, ihre Position in der Artikelhierarchie zu bestimmen. Die „1)“ kann die erste Gliederungsebene einleiten, ebenso gut aber auch die zweite oder dritte Ebene bezeichnen, je nachdem, ob der Artikel mit „I. A. 1)“ oder „1) a)“ beginnt. Auch die Abfolge der Gliederungsmarken variiert. Im Artikel *herz* beispielsweise wird die oberste Gliederungsebene durch Großbuchstaben eingeleitet, auf der nächstniedrigeren Ebene gefolgt von römischen Ziffern („A. I. 1)“); im Artikel *gott* dagegen kennzeichnen römische Ziffern die oberste, Großbuchstaben die zweite und arabische Ziffern die dritte Gliederungsebene („I. A. 1)“). Je differenzierter und tiefer die Artikel untergliedert sind, desto mehr häufen sich die Besonderheiten.[22] Das genaue Erfassen der Position einer Gliederungsmarke in der Gliederungshierarchie stellte daher eine große Herausforderung bei der Programmierung der Markup-Routinen dar.

Ähnlich problematisch gestaltete sich die Auszeichnung der Quellensiglen in einigen der retrodigitalisierten Wörterbücher. Im BMZ beispielsweise wird Lorenz Diefenbachs *Vergleichendes Wörterbuch der gotischen Sprache* in mindestens zwölf Varianten angeführt. Es begegnet als „*Diefenb. g. wb.*“, „*Diefenb. g. w.*“, „*Diefenb. g. wtrb.*“, „*Diefenb. g. wrtbch.*“, „*Diefenb. g. wtrbch.*“, „*Diefenb. g. wörterb.*“, „*Diefenb. goth. w.*“, „*Diefenb. goth. wb.*“, „*Diefenb. goth. wtrbch.*“, „*Diefenb. goth. wörterb.*“, „*Diefenbach g. wb.*“, „*Diefenbach goth. wtrbch*.“ und „*Diefenbach goth. wörterb.*“ Versehentlich fehlende Abkürzungspunkte erhöhen die Varianz, die sich auch durch Zuhilfenahme des Quellenverzeichnisses nicht auflösen lässt, da dessen Zusammensteller Nellmann nur die erste der angeführten Siglen aufgenommen hat.[23] Ähnlich ist der die Siglen betreffende Befund im [1]DWb. Abraham a Santa Clara etwa wird als „ABRAHAM A S. CL.“, „ABRAHAM VON S. CLARA“, „ABRAH. A SANTA CLARA“, „ABR. A ST. CLARA“, „ABR. A. S. CL.“, „ABR. A. S. C.“ usw. angeführt. Das Quellenverzeichnis führt von den mehr als zwanzig Varianten lediglich eine auf. Um dem Nutzer des digitalen

[22] Zur Problematik der Gliederungsmarken im [1]DWb vgl. Bartz/Burch/Christmann/Gärtner/Hildenbrandt/Schares/Wegge (2004), S. 81 f.

[23] Vgl. dazu Burch/Fournier (2001), S. 143.

[1]DWb die erfolgreiche Suche nach „Abraham a Santa Clara“ zu ermöglichen, mussten alle Varianten ausfindig gemacht und auf die normalisierte Form abgebildet werden.

Als sehr aufwendig erwies sich bei einigen Wörterbüchern auch die Auszeichnung der Binnenverweise. Sie als Hyperlinks zu realisieren, gehört eigentlich zu den Mindestanforderungen einer Internetversion, im PfälzWb und im [1]DWb beispielsweise erwies sich die dafür notwendige Datenaufbereitung allerdings keineswegs als trivial. Denn wo sich das GWb mit einem Verweispfeil, einem „vgl“ oder einem „Syn“ zur Einleitung eines Verweises auf ein anderes Stichwort begnügt, wurden für das PfälzWb an die hundert verschiedene Verweiseinleitungen von „vgl.“ und „s.“ über „dafür“ und „hierfür“ bis hin zu „im Unterschied zu“ und „im Ggs. zu“ ermittelt. Das [1]DWb überbietet diesen Befund noch, indem es neben dem gängigen „s.“ oder „vgl.“ in allen Variationen („vgl. oben“, „vergleiche“, „vergl. auch“, „s.“, „siehe“, „s. d. und vgl. unten“) zahllose diskursive Wendungen wie beispielsweise „früher hiesz es auch“, „man sehe die ausführung unter“ oder „wofür man doch lieber sagt“ als Verweiseinleitungen nutzt. Ein Ermitteln aller Verweiseinleitungen erscheint hier ebenso unmöglich wie die klare Abgrenzung der Verweislemmata, die häufig nur unzureichend durch typographische oder anderweitige Strukturmarker von den umgebenden Informationen geschieden sind. Vollständigkeit ist in der Auszeichnung der Binnenverweise daher bei einem Wörterbuch mit dem Umfang und der Anlage des [1]DWb selbst durch manuelle Nacharbeit nicht zu erreichen.

Auch eine Vernetzung verschiedener lexikographischer Ressourcen auf Lemmaebene stellt eine Herausforderung dar in den Fällen, in denen die Lemmata der einzelnen Wörterbücher nach verschiedenen Grundsätzen angesetzt sind oder sehr viele homographe Lemmata vorkommen. Solche Probleme wurden virulent bei der Verlinkung des GWb mit dem [1]DWb. Das Phänomen einer abweichenden Typographie der Lemmata in beiden Wörterbüchern konnte teilweise durch eine „Normalisierung“ der Stichwörter abgefangen werden, das heißt, Umlaute wurden auf die Grundbuchstaben reduziert, Akzente eliminiert usw. Dadurch konnten bei einem automatischen Vergleich der Lemmalisten von GWb und [1]DWb eine Reihe von Übereinstimmungen und somit Verweisziele ermittelt und es konnte zum Beispiel das Stichwort *abbüßen* im GWb mit dem Stichwort *abbüszen* im [1]DWb verlinkt werden. Bei homographen Stichwörtern dagegen konnte die Zuordnung nicht automatisch erfolgen. Hier musste ein verstehender Leser die mittels automatisierter Verfahren eingetragenen Verweise in einem manuellen Nachbearbeitungsschritt kontrollieren und gegebenenfalls korrigieren, um beispielsweise zu verhindern, dass aus dem GWb-Artikel *Buch* in den [1]DWb-Artikel *buch* im Sinne von „buk, *backte*“ ([1]DWb, Band 2, Spalte 466) verlinkt wurde.

Nahezu komplett versagen muss der Versuch eines maschinellen Auszeichnungsverfahrens bei nur implizit gegebenen Informationen. Sich häufig auf Belegstellenangaben beziehende Rückverweise wie „ebend.“, „ebd.“, „das.“, „ebendas.“ lassen sich von „einem ‚kontextsensitiven‘ menschlichen Bearbeiter“ (Burch/Fournier 2001, S. 144) erschließen. Dieses kontextsensitive Erschließen des Verweisziels in eine Auszeichnungsroutine zu fassen, ist dagegen ein schwieriges, zum Teil sicherlich unlösbares Unterfangen. Unter Punkt 10) des Artikels *gericht* beispielsweise führt das [1]DWb einen mehrfachen Rückverweis auf: „gericht *am wagen, lateralia* SCHMELLER a. a. o.; ingricht, *vorrichtung am wagen. ebd. aus dem bairischen wald*.“ Herauszufinden, auf welches Werk Schmellers sich „*ebd.*“ bezieht, ist selbst für den menschlichen Benutzer nicht ganz einfach, denn der nächste Hinweis auf Schmeller findet sich in der gedruckten Fassung gut vierzig Zeilen über dem angeführten Zitat und dort leider ebenfalls mit der Angabe „a. a. o.“. Dies wiederholt sich zehn Zeilen zuvor, so dass im Artikel *gericht* insgesamt zwar viermal auf Schmeller verwiesen wird, aber jedes Mal ohne Nennung des Werktitels. Ein kundiger Wörterbuchnutzer vermag zu erschließen, dass es sich

hierbei um Schmellers *Bayerisches Wörterbuch* (Band 2/1 [1877], Spalte 35 ff., Artikel *richten*) handeln muss, einem maschinellen Markup sind hier jedoch definitiv Grenzen gesetzt.

4. Resümee

Insgesamt hat sich die Entscheidung, bei der Auszeichnung der retrodigitalisierten Wörterbücher auf die TEI-Guidelines zu setzen, bewährt. Sicherlich eignen sie sich nicht zuletzt deshalb für die Codierung ursprünglich nur gedruckt erschienener lexikographischer Ressourcen, weil diese die Ausgangsbasis für die Entwicklung des Wörterbuchmoduls waren. Auch dass die Guidelines bewusst allgemein gehalten sind – ein Umstand, der andere, digital entstehende Vorhaben von einer TEI-basierten Modellierung Abstand nehmen lässt[24] –, erweist sich im Falle der retrodigitalisierten Wörterbücher als Vorzug. Probleme bei der TEI-konformen Auszeichnung dieser Wörterbücher liegen weniger im Codierungsschema begründet als vielmehr in der Anlage der Wörterbücher. Diskursivität und geringe Standardisierung von Wörterbuchartikeln, das Fehlen eindeutiger struktureller Marker für bestimmte lexikographische Informationseinheiten und Varianz sind Phänomene, die sich mittels computergestützter Verfahren nicht abfangen lassen. Nicht die Anforderungen der potentiellen Nutzer an ein elektronisches Wörterbuch und die Funktionalitäten des Computers entscheiden in diesen Fällen über die Art und Tiefe der Textauszeichnung, sondern vor allem die Beschaffenheit des gedruckten Ausgangsmaterials und die für dessen Codierung zur Verfügung stehenden personellen und finanziellen Ressourcen. Das „Wunschziel" einer „linguistisch motivierte[n] und feinkörnige[n] Modellierung" würde – darin ist Storrer voll und ganz zuzustimmen –, „wenn sie auf der Grundlage eines gedruckten Wörterbuchs erfolgt, einen relativ hohen Auf- und Nachbearbeitungsaufwand" (Storrer 2001, S. 64) erfordern. Für ein Wörterbuch vom Umfang des [1]DWb beispielsweise ist jedoch eine solche Auf- und Nachbearbeitung gänzlich undenkbar und unfinanzierbar. Dennoch sollte auf eine Retrodigitalisierung und TEI-konforme Auszeichnung nicht aus dem Grund verzichtet werden, dass der Datenmodellierung Grenzen gesetzt sind. Denn auch wenn unter Umständen keine hohe Auszeichnungstiefe erreicht werden kann, bieten sie doch andere Zugriffs- und Auswertungsmöglichkeiten als ihre gedruckten Pendants,[25] Zugriffs- und Auswertungsmöglichkeiten, die – das belegen die Nutzerzahlen[26] – gerne genutzt werden.

5. Literatur

5.1 Wörterbücher im Trierer Wörterbuchnetz

Adelung, Johann Christoph (1793-1801): Grammatisch-Kritisches Wörterbuch der Hochdeutschen Mundart mit beständiger Vergleichung der übrigen Mundarten, besonders aber der oberdeutschen. Zweyte, vermehrte und verbesserte Ausgabe. 4 Bände. Leipzig.

BMZ = Benecke, Georg Friedrich/Müller, Wilhelm/Zarncke, Friedrich (1854-1866): Mittelhochdeutsches Wörterbuch. Leipzig. [Nachdruck mit einem Vorwort und einem zusammengefaßten Quellenverzeichnis von Eberhard Nellmann sowie einem alphabetischen Index von Erwin Koller, Werner Wegstein und Norbert Richard Wolf. Stuttgart 1990.]

[1]DWb = Grimm, Jacob und Wilhelm (1854-1960. 1971): Deutsches Wörterbuch. 16 Bde. in 32 Teilbänden. Nebst Quellenverzeichnis. Leipzig.

[24] Im Projekt *elexiko* beispielsweise entschied man sich aus diesem Grund gegen die von der TEI vorgeschlagene Standard-Modellierung von Wörterbüchern. Vgl. Müller-Spitzer (2005), S. 28.

[25] S. dazu oben Anm. 13.

[26] Monatlich sind rund eine Million Zugriffe auf das Trierer Wörterbuchnetz zu verzeichnen.

ElsWb = Martin, Ernst/Lienhart, Hans (1899-1907): Wörterbuch der elsässischen Mundarten. 2 Bände. Straßburg. [Ndr.: Berlin/New York 1974.]

Findebuch = Gärtner, Kurt/Gerhardt, Christoph/Jaehrling, Jürgen/Röll, Walter/Timm, Erika/Hanrieder, Christoph (Datenverarbeitung) (1992): Findebuch zum mittelhochdeutschen Wortschatz. Mit einem rückläufigen Index. Stuttgart.

GWb = Goethe-Wörterbuch (1966 ff.). Hg. v. der Berlin-Brandenburgischen Akademie der Wissenschaften, der Akademie der Wissenschaften zu Göttingen und der Heidelberger Akademie der Wissenschaften. Band 1 ff. Stuttgart.

Lexer = Lexer, Matthias (1872-1878): Mittelhochdeutsches Handwörterbuch. Leipzig. [Nachdruck mit einer Einleitung von Kurt Gärtner. Leipzig 1992.]

LothrWb = Follmann, Michael Ferdinand (1909): Wörterbuch der deutsch-lothringischen Mundarten. Leipzig 1909. [Ndr.: Hildesheim/New York 1971.]

Meyers Großes Konversationslexikon. Ein Nachschlagewerk des allgemeinen Wissens. Sechste, gänzlich neubearbeitete und vermehrte Auflage. Leipzig und Wien 1905-1909.

PfälzWb = Pfälzisches Wörterbuch. Begründet von Ernst Christmann, fortgeführt von Julius Krämer, bearbeitet von Rudolf Post unter Mitarbeit von Josef Schwing. 6 Bände. Wiesbaden/Stuttgart 1965-1997.

RheinWb = Rheinisches Wörterbuch. Im Auftrag der Preußischen Akademie der Wissenschaften, der Gesellschaft für Rheinische Geschichtskunde und des Provinzialverbandes der Rheinprovinz auf Grund der von Johannes Franck begonnenen, von allen Kreisen des Rheinischen Volkes unterstützten Sammlung bearbeitet und herausgegeben von †Josef Müller, †Heinrich Dittmaier, Rudolf Schützeichel und Matthias Zender. 9 Bände. Bonn/Berlin 1928-1971.

5.2 Forschungsliteratur

Bartz, Hans-Werner/Burch, Thomas/Christmann, Ruth/Gärtner, Kurt/Hildenbrandt, Vera/Schares, Thomas/Wegge, Klaudia (2004): Wie das Deutsche Wörterbuch in den Computer kam. In: Der Digitale Grimm. Deutsches Wörterbuch von Jacob und Wilhelm Grimm. Elektronische Ausgabe der Erstbearbeitung. Bearbeitet von Hans-Werner Bartz, Thomas Burch, Ruth Christmann, Kurt Gärtner, Vera Hildenbrandt, Thomas Schares, Klaudia Wegge. Hg. vom Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier in Verbindung mit der Berlin-Brandenburgischen Akademie der Wissenschaften. 2 CD-ROMs, Benutzerhandbuch, Begleitbuch. Frankfurt/Main, S. 73-90.

Burch, Thomas/Fournier, Johannes (2001): Zur Anwendung der TEI-Richtlinien bei der Retrodigitalisierung mittelhochdeutscher Wörterbücher. In: Lemberg, Ingrid/Schröder, Bernhard/Storrer, Angelika (Hg.): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher (= Lexicographica: Series maior; 107). Tübingen, S. 133-153.

Christmann, Ruth/Hildenbrandt, Vera/Schares, Thomas (2001): Ein „heiligthum der sprache“ digitalisiert: Das Deutsche Wörterbuch von Jacob und Wilhelm Grimm auf CD-ROM und im Internet. In: Benito, Nicolás Castrillo/Stahl, Peter (Hg.): TUSTEP educa. Actas de Congreso del International Tustep User Group: Peñaranda de Duero (Burgos) Octubre 1999. Burgos, S. 13-35.

Burnard, Lou/Bauman, Syd (Hg.) (2010): TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford, Providence, Charlottesville, Nancy. Internet: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf. (Stand: November 2011).

Gloning, Thomas/Schlaps, Christiane (1999): Prototypen für ein elektronisches Goethe-Wörterbuch. In: Sprache und Datenverarbeitung. International Journal for Language Data Processing 32, 2, S. 21-34.

Gloning, Thomas/Welter, Rüdiger (2001): Wortschatzarchitektur und elektronische Wörterbücher: Goethes Wortschatz und das Goethe-Wörterbuch. In: Lemberg, Ingrid/Schröder, Bernhard/Storrer, Angelika (Hg.): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. (= Lexicographica: Series maior; 107). Tübingen, S. 117-132.

Ide, Nancy/Véronis, Jean (1996a): Encodage des dictionnaires électroniques: problèmes et propositions de la TEI. In: Piotrowski, David (Hg.): Lexicographie et informatique. Autour de l’informatisation du Trésor de la Langue Française. Actes du Colloque International de Nancy. Paris, S. 239-261. Internet: http://sites.univ-provence.fr/~veronis/pdf/1996Tlf.pdf. (Stand: November 2011).

Ide, Nancy/Véronis, Jean (1996b): Codage TEI des dictionnaires électroniques. In: Cahiers GUTenberg n° 24 (spécial TEI), S. 170-176.

Jannidis, Fotis (1997): Wider das Altern elektronischer Texte: philologische Textauszeichnung mit TEI. In: editio. Internationales Jahrbuch für Editionswissenschaft 11, S. 152-177.

Müller-Spitzer, Carolin (2005): Die Modellierung lexikographischer Daten und ihre Rolle im lexikographischen Prozess. In: Haß, Ulrike (Hg.): Grundfragen der elektronischen Lexikographie. elexiko – das Online-Informationssystem zum deutschen Wortschatz. (Schriften des Instituts für Deutsche Sprache; 12) Berlin/New York, S. 21-54.

Rennie, Susan (2000): Encoding a Historical Dictionary with the TEI. (With reference to the Electronic Scottish National Dictionary Project). In: Heid, Ulrich/Evert, Stefan/Lehmann, Egbert/Rohrer, Christian (Hg.): Proceedings of the Ninth EURALEX International Congress, EURALEX 2000. Stuttgart, S. 261-271. Internet: http://www.euralex.org/elx_proceedings/Euralex2000/031_Susan%20RENNIE_Encoding%20a%20Historical%20Dictionary%20with%20the%20TEI%20(With%20reference%20to%20the%20Electronic%20Scottish%20National%20Dictionary%20Project).pdf. (Stand: November 2011).

Schmidt, Frieder (1997): Neuland für die Buchgeschichte – Quellenaufbereitung im Zeitalter des WWW. Hypertext Markup Language (HTML), Standard Generalized Markup Language (SGML) und die Guidelines for Electronic Text Encoding and Interchange der Text Encoding Initiative (TEI). In: Leipziger Jahrbuch zur Buchgeschichte 7, S. 343-365.

Schmidt, Hartmut (2004): Das Deutsche Wörterbuch. Gebrauchsanleitung. In: Der Digitale Grimm. Deutsches Wörterbuch von Jacob und Wilhelm Grimm. Elektronische Ausgabe der Erstbearbeitung. Bearbeitet von Hans-Werner Bartz, Thomas Burch, Ruth Christmann, Kurt Gärtner, Vera Hildenbrandt, Thomas Schares, Klaudia Wegge. Hg. vom Kompetenzzentrum für elektronische Erschließungs- und Publikationsverfahren in den Geisteswissenschaften an der Universität Trier in Verbindung mit der Berlin-Brandenburgischen Akademie der Wissenschaften. 2 CD-ROMs, Benutzerhandbuch, Begleitbuch. Frankfurt/Main, S. 25-64.

Schmidt, Hartmut/Reinitzer, Heimo/Kühlmann, Wilhelm (2004): Vorwort. In: Goethe-Wörterbuch. Vierter Band, Lieferung 12. Stuttgart.

Storrer, Angelika (2001): Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: Lemberg, Ingrid/Schröder, Bernhard/Storrer, Angelika (Hg.): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. (= Lexicographica: Series maior; 107). Tübingen, S. 53-69.

Der Aufbau einer maßgeschneiderten XML-basierten Modellierung für ein Wörterbuchnetz

Carolin Müller-Spitzer mueller-spitzer@ids-mannheim.de, Tel.: +49 621 1581-429

1. Einleitung

Im vorliegenden Beitrag soll der Aufbau einer maßgeschneiderten XML-Modellierung für ein Wörterbuchnetz erläutert werden. Diese Schriftfassung beruht auf einem gleichlautenden Vortrag, der auf dem ersten Arbeitstreffen des DFG-Netzwerks „Internetlexikografie“ in Mannheim im Mai 2011 gehalten wurde. Der Beitrag ist als Werkstattbericht zu verstehen, d. h. als praktisch orientierter Blick sowohl darauf, wie wir unsere Modellierung für OWID konzipiert haben, welche Konsequenzen dies für die lexikographische Arbeit sowie für die Recherchemöglichkeiten der Nutzer hat, als auch darauf, welche Vor- und Nachteile wir bei diesem Modellierungsansatz sehen. Der vorliegende Beitrag bietet damit keine umfassende theoretische Auseinandersetzung mit verschiedenen Möglichkeiten der Modellierung. Lediglich im folgenden Kapitel werden die Grundzüge des Modellierungsansatzes kurz erläutert und es wird auf entsprechende weiterführende projektbezogene Literatur verwiesen.

2. Allgemeines zur Modellierung in OWID

Das Online-Wortschatz-Informationssystem Deutsch (OWID) ist das Wörterbuchportal des Instituts für Deutsche Sprache (IDS) in Mannheim (s. www.owid.de). Es beinhaltet wissenschaftliche, korpusbasierte Wörterbücher zum Deutschen mit unterschiedlichen inhaltlichen Schwerpunkten. Dies sind im Moment[27]:

- *elexiko*: *elexiko* verfügt über eine umfangreiche korpusbasiert gewonnene Stichwortliste zum Deutschen mit über 300.000 Einträgen. Zu fast allen Stichwörtern sind in *elexiko* automatisch gewonnene Textbelege sowie orthographische Angaben zu finden. Darüber hinaus sind über 1.500 hochfrequente Stichwörter im „Lexikon zum öffentlichen Sprachgebrauch“ ausführlich lexikographisch beschrieben.
- Feste Wortverbindungen: In diesem Bereich sind lexikographische Ergebnisse der korpusgesteuerten Mehrwortforschung veröffentlicht. Die Wortartikel haben unterschiedliche linguistische Beschreibungstiefen und Darstellungsformate. Derzeit sind etwa 130 Mehrwortartikel in OWID enthalten (weiterführende Informationen s. Wortverbindungen online).
- Neologismenwörterbuch: Das Neologismenwörterbuch präsentiert in über 1.000 Wortartikeln neue Wörter bzw. Wortverbindungen sowie neue Bedeutungen von etablierten Wörtern, die in den 90er Jahren des 20. Jahrhunderts in die Allgemeinsprache eingegangen sind.
- Schulddiskurs 1945-55: In diesem Wörterbuch sind 85 Haupt- sowie über 200 Unterstichwörter zum Schulddiskurs im ersten Nachkriegsjahrzehnt verzeichnet. Dieser Wortschatzbereich ist aus einem breit angelegten Korpus von Texten, die in den Jahren 1945-55 erschienen sind, erarbeitet worden.

[27] Für projektbezogene Literatur s. die Projektseiten unter www.ids-mannheim.de, dort jeweils den Punkt „Publikationen“. Alle genannten Wörterbücher haben außerdem unter OWID eigene Begleittexte.

Neben Wörterbüchern enthält OWID eine Online-Bibliographie zur elektronischen Lexikographie (OBELEX) sowie eine Datenbank zu Online-Wörterbüchern.

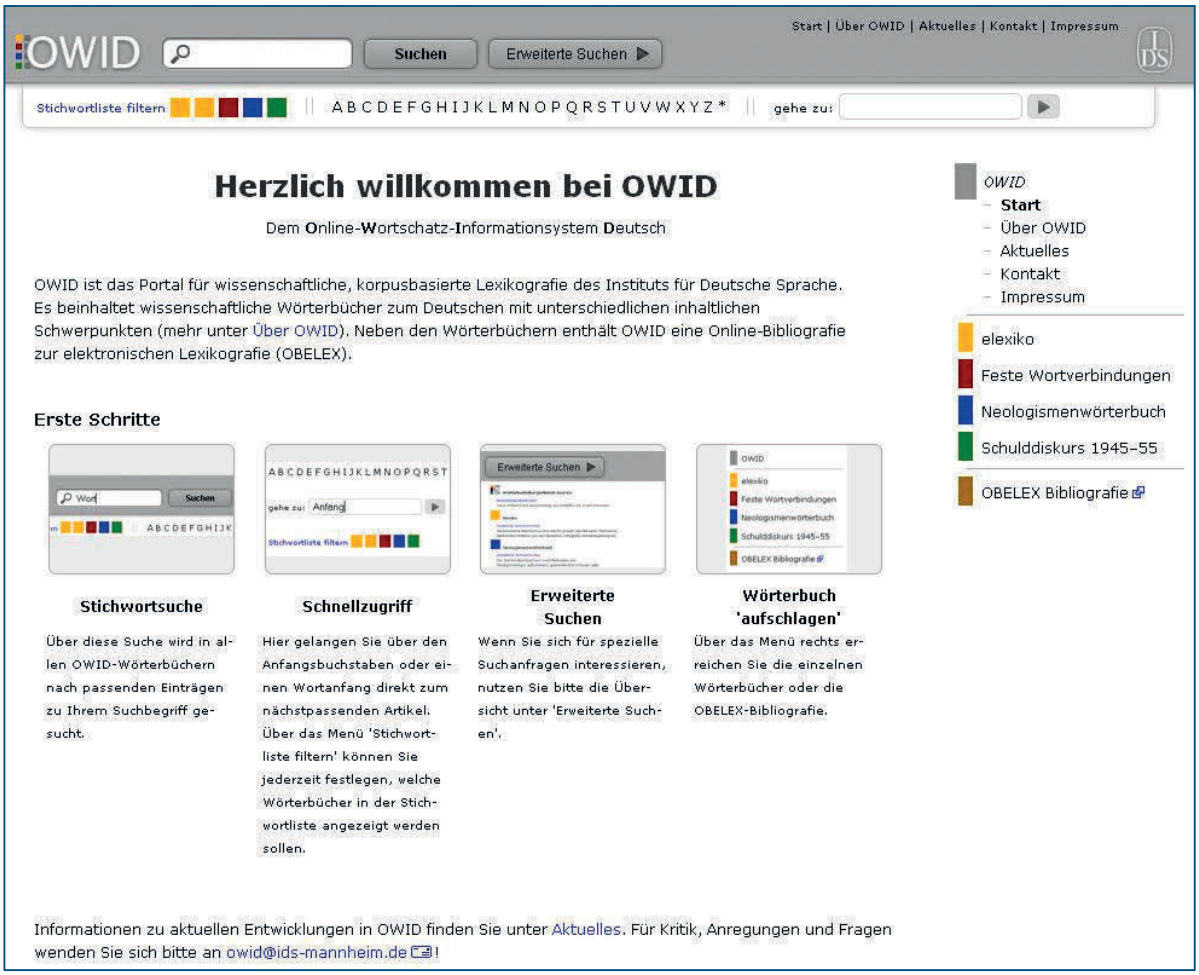


Abb. 1: Startseite von OWID

OWID wird kontinuierlich erweitert. Im Moment wird an der Integration des „E-ValBU“ (http://hypermedia2.ids-mannheim.de/evalbu/index.html), des „Handbuchs deutscher Kommunikationsverben“ (Harras et al. 2004), der „Schlüsselwörter der Wendezeit“ (Herberg et al. 1997) und eines Sprichwörterbuchs gearbeitet. Bei Letzterem handelt es sich um 300 Artikel zu deutschen Sprichwörtern, die im Rahmen des EU-Projekts SprichWort[28] erarbeitet wurden. Alle diese Wörterbücher werden gezielt für eine Publikation im Internet aufbereitet, sowohl was die Art der Datenstrukturierung als auch was die Form der Darstellung betrifft.

OWID hat seine Arbeit als eigenes Projekt erst 2008 aufgenommen. Vorgänger war jedoch das *elexiko*-Portal, welches schon sechs Jahre früher begonnen wurde (vgl. Müller-Spitzer 2008a, b; Klosa 2008). Die Modellierung, die allen Wörterbüchern von OWID zu Grunde liegt, wurde daher bereits in den Jahren 2002ff. entwickelt. Die Leitlinien, die der Modellierung zu Grunde liegen, sind folgende (für eine ausführliche Begründung und Erläuterung der Richtlinien s. Müller-Spitzer 2007a, 2008a, 2008b, 2011):

[28] Siehe dazu auch die IDS-Projektbeschreibung. Die SprichWort-Artikel sind ebenso auf der Sprichwortplattform, in der Sprichwortdatenbank Deutsch, abrufbar (http://www.sprichwort-plattform.org/sp/Sprichwort).

- Die Modellierung ist XML-basiert, um die Softwareunabhängigkeit und Langlebigkeit der Daten zu gewährleisten.
- Die Modellierung ist maßgeschneidert, um einen genauen Zuschnitt auf die Erfordernisse der einzelnen Wörterbücher im Portal sicherzustellen.
- Die Modellierung ist sehr ‚streng' und genau, um die Lexikographen bei der Einhaltung der formalen Artikelstruktur so gut wie möglich zu unterstützen.
- Die Modellierung ist so granular wie möglich, um eine bestmögliche Flexibilität hinsichtlich der Darstellung und des Zugriffs zu gewährleisten. Diese letzten beiden Aspekte können mit dem Terminus *Inhaltsstrukturmodellierung* zusammengefasst werden (vgl. Müller-Spitzer 2007a, 152ff.).

Um Objekte, die in verschiedenen Wörterbüchern in OWID vorkommen, nur einmal zu modellieren, wurde eine DTD-Bibliothek für OWID angelegt.[29] In dieser DTD-Bibliothek finden sich DTDs mit Bausteinen, die in allen Wörterbüchern vorkommen (wie Belege, Kommentare, Abbildungen etc.), sowie DTDs für objektübergreifende Gruppen (wie Elemente, die sowohl für die Einwortlemmata in *elexiko* sowie für die Neologismen benötigt werden), Bausteinen für einzelne Wörterbücher (z.B. für die Neologismen der Neologismentyp, der sowohl bei Einwort- wie bei Mehrwortlemmata angegeben wird) und zuletzt die Kopf-DTDs für die einzelnen Wörterbücher, die aus der DTD-Bibliothek die Elemente zusammenziehen, die für das jeweilige Wörterbuch relevant sind. So kann man am Aufbau der DTD-Bibliothek bereits erkennen, welche Angaben wörterbuchübergreifend gleich modelliert sind und sich daher z.B. für erweiterte, wörterbuchübergreifende Suchen eignen. Außerdem wird auch deutlich, dass es sich bei OWID nicht um ein Wörterbuchportal mit völlig unterschiedlichen, untereinander nicht verbundenen Ressourcen handelt, sondern um ein Wörterbuchnetz (im Sinne von Engelberg/Müller-Spitzer, im Erscheinen). Im Folgenden sollen die Modellierungsprinzipien anhand eines Beispiels – der Angaben zur Wortbildung – illustriert werden.

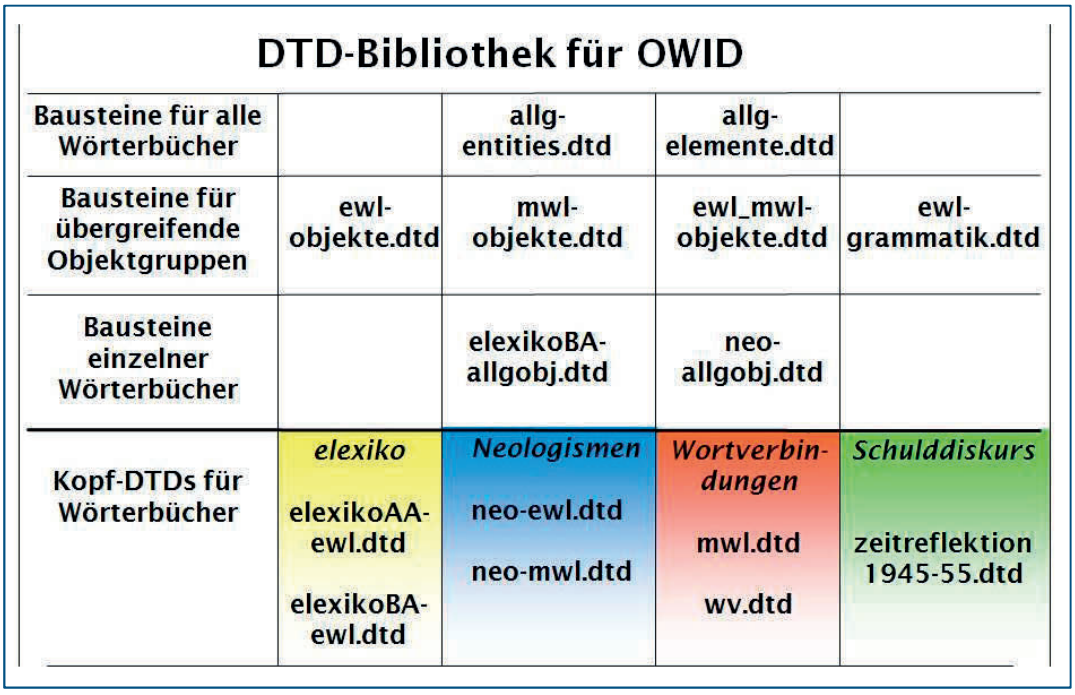


Abb. 2: DTD-Bibliothek für OWID

[29] Da die Modellierung bereits 2002 entwickelt wurde, wurden XML-DTDs verwendet. Intern werden diese mittlerweile zu XML-Schemata konvertiert.

3. Beispiel: Angaben zur Wortbildung

3.1 Modellierung

Die Angaben zur Wortbildung, die für *elexiko* und das Neologismenwörterbuch (für die Einwortlemmata) einheitlich sind, sind analog zum Modellierungsansatz sehr granular ausgezeichnet. Es werden nicht nur die Wortbildungsarten wie Ableitung, Zusammensetzung etc. codiert, sondern es werden auch die einzelnen gebildeten Teile näher bestimmt. Bei Präverbfügungen wie „weißwaschen“ (*elexiko*) oder „schönrechnen“ (Neologismenwörterbuch) (vgl. Abbildung 3 und 4) werden beispielsweise nicht nur die Art der Wortbildung (Präverbfügung), sondern auch die einzelnen Bestandteile (Präverb, verbale Basis) einzeln codiert und – in *elexiko* – die Wortbildungsbedeutung spezifiziert. Die einzelnen Bestandteile werden – wenn möglich – mit den korrespondierenden Wortartikeln in der Datenbasis verlinkt (ref-ID-Attribute). Online wird nur ein kleiner Teil dieser in der XML-Instanz codierten Informationen angezeigt (vgl. Klosa 2011).

```
<vb-wortbildung>
<praeverbfg>
<praeverbA
     basistyp="adjektiv"
     artikel-refid="137003"
     lesart-refid="0">weiß</praeverbA>
<verb-basisA
     basistyp="verb"
     artikel-refid="136664"
     lesart-refid="0">waschen</verb-basisA>
    </praeverbfg>
<vb-wortblgbedeutungA
     bezeichnung="aktive-zustandsveraenderg"/>
</vb-wortbildung>
```

Abb. 3: Ausschnitt aus der XML-Instanz zum Wortartikel weißwaschen (*elexiko*)

```
<vb-wortbildung>
<praeverbfg>
     <praeverbA
     basistyp="adjektiv"
     artikel-refid="288474"
     lesart-refid="0">schön</praeverbA>
<verb-basisA
     basistyp="verb"
     artikel-refid="283509"
     lesart-refid="0">rechnen</verb-basisA>
</praeverbfg>
</vb-wortbildung>
```

Abb. 4: Ausschnitt aus der XML-Instanz zum Wortartikel schönrechnen (Neologismenwörterbuch)

3.2 Recherchemöglichkeiten für die Lexikographen

Wichtige Gründe, weshalb die Modellierung bzw. Auszeichnung der lexikographischen Daten in OWID so feingranular ist, ist zum einen die Unterstützung der Lexikographen bei der Einhaltung der formalen Artikelstruktur, aber genauso sind es auch die Möglichkeiten, sehr gezielt auf die erarbeiteten Daten zugreifen zu können. So können z.B. Konsistenzprüfungen über verschiedene Wortartikel hinweg erheblich erleichtert werden.

Alle Daten von OWID werden im „Electronic Database Administration System (EDAS)“ gespeichert, einem Datenbankmanagementsystem basierend auf Oracle 11.[30] Intern können die beteiligten Lexikographen über OWID eine XPath-Suche benutzen, in der die Autoren alles, was XML-basiert ausgezeichnet ist, auch abfragen können. So kann beispielsweise im Bereich der Wortbildung nach allen Präverbfügungen gesucht werden (`//praeverbfg`, vgl. Abbildung 5) oder auch spezifischer nach allen Präverbfügungen mit einem Adjektiv als Präverb (`//praeverbA[@basistyp="adjektiv"]`, vgl. Abbildung 6). Genauso kann auch auf alle Lemmata, bei denen eine bestimmte Wortbildungsbedeutung angegeben wurde, zugegriffen werden.

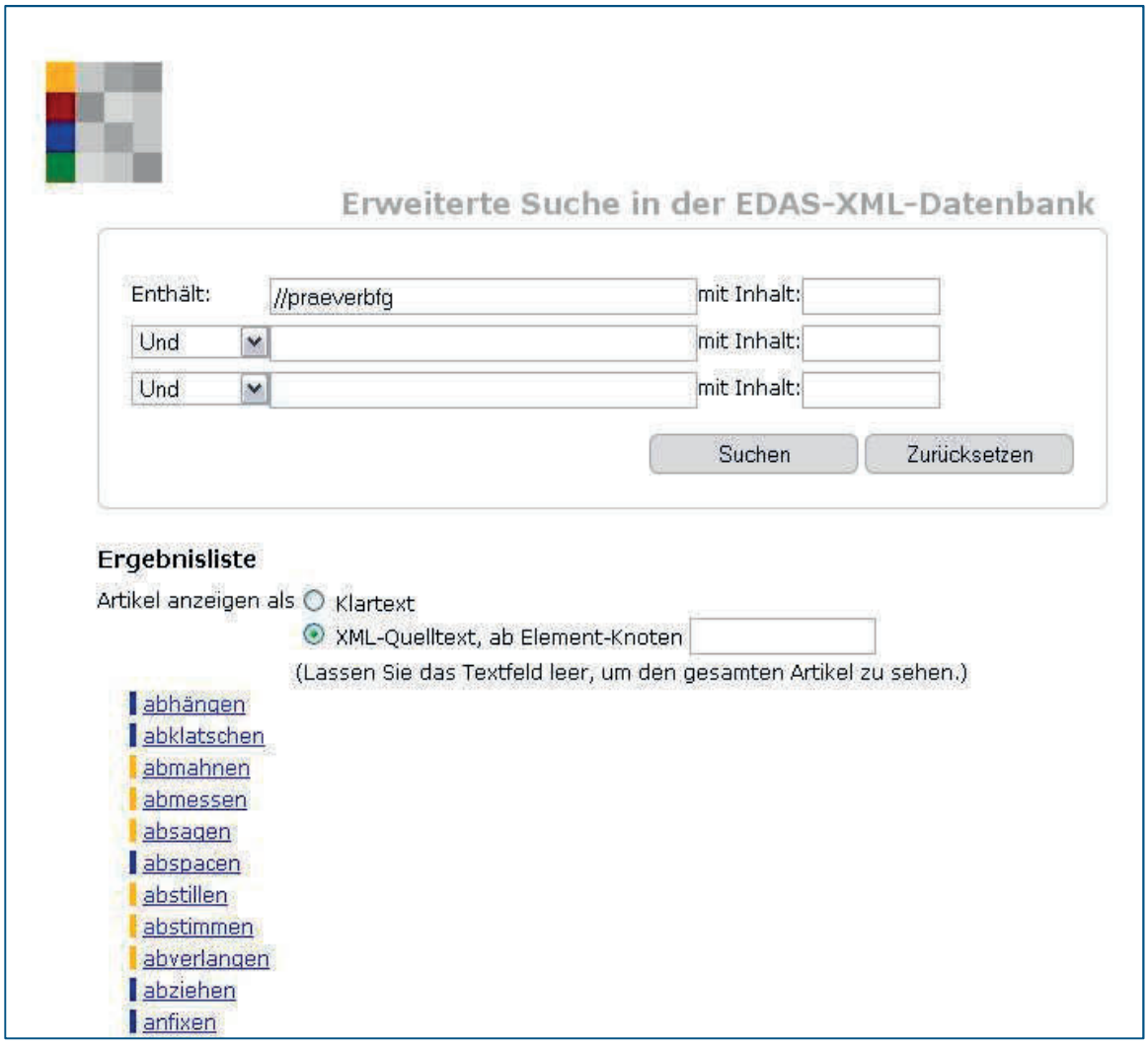


Abb. 5: XPath-basierte interne Suche über OWID nach allen Präverbfügungen

[30] EDAS wurde von Roman Schneider aus der Abteilung Grammatik des IDS entwickelt.

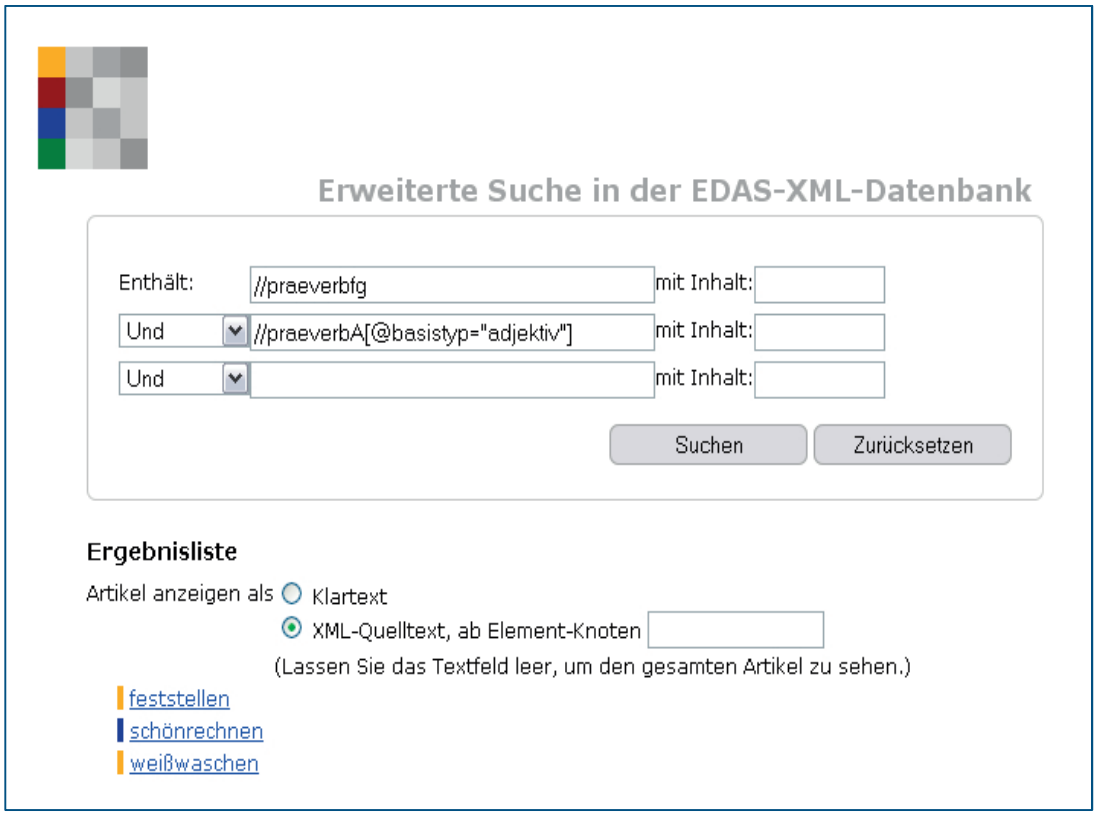


Abb. 6: XPath-basierte interne Suche über OWID nach allen Präverbfügungen mit einem adjektivischen Präverb

3.3 Erweiterte Suchen für die Nutzer

Für Endbenutzer eignen sich diese XPath-basierten Suchen nicht, denn Voraussetzung für eine erfolgreiche Benutzung ist die genaue Kenntnis der XML-Modellierung. Dies würde bedeuten, dass – im Falle von *elexiko* – ein Benutzer über 400 Elemente und über 300 zugehörige Attribute kennen müsste, d.h., die vollständige XML-Struktur müsste nach außen so dokumentiert sein, dass Außenstehende sich einarbeiten könnten. Zusätzlich müsste die XPath-Syntax erlernt werden. Die Hürden, eine solche Suche benutzen zu können, sind demnach zu hoch. Außerdem eröffnet eine solche Suche einen zu offenen Zugriff auf die lexikographischen Daten, der vor Missbrauch (beispielsweise dem Herunterladen ganzer Artikel) schwer zu schützen ist.

Für die Benutzer haben wir deshalb eine andere Form der erweiterten Suche für den Bereich der Wortbildung entwickelt, mit der wir versucht haben, die Gliederung des Angabebereichs graphisch zu veranschaulichen und so den Benutzern einen leichten Zugang zu sehr spezialisierten Suchen zu eröffnen. Diese neue Form der erweiterten Suche ist noch im Pilotstadium und soll voraussichtlich Ende 2011 unter den erweiterten Suchen von OWID online verfügbar sein (vgl. Abbildung 7 und 8).

Die Abbildung kann den interaktiven Aufbau der graphischen Suche nur bedingt verdeutlichen, deshalb wird das Vorgehen kurz erläutert: Klickt ein Benutzer auf den Kasten „Wortbildungsart“, werden folgende Kästen expandiert: „Derivation“, „Komposition“, „Präverbfügung“ und „Kurzwortbildung“. Klickt man wiederum auf eine dieser Arten, expandiert ggf. der nächste relevante Teil des Strukturbaumes; je nach Angabe wiederum als Baum oder als Auswahlmenü (wie im Bereich der Präverbfügung, s. Abbildung 7). Ist schon durch die Aus-

wahl der Wortbildungsart das Suchergebnis hinreichend überschaubar, wird ein kurzer erläuternder Dialog angezeigt (wie im Bereich der Kontamination, vgl. Abbildung 8). Das Suchergebnis enthält bei dieser erweiterten Suche bereits alle relevanten Angaben als Auszug aus den zugehörigen XML-Instanzen.

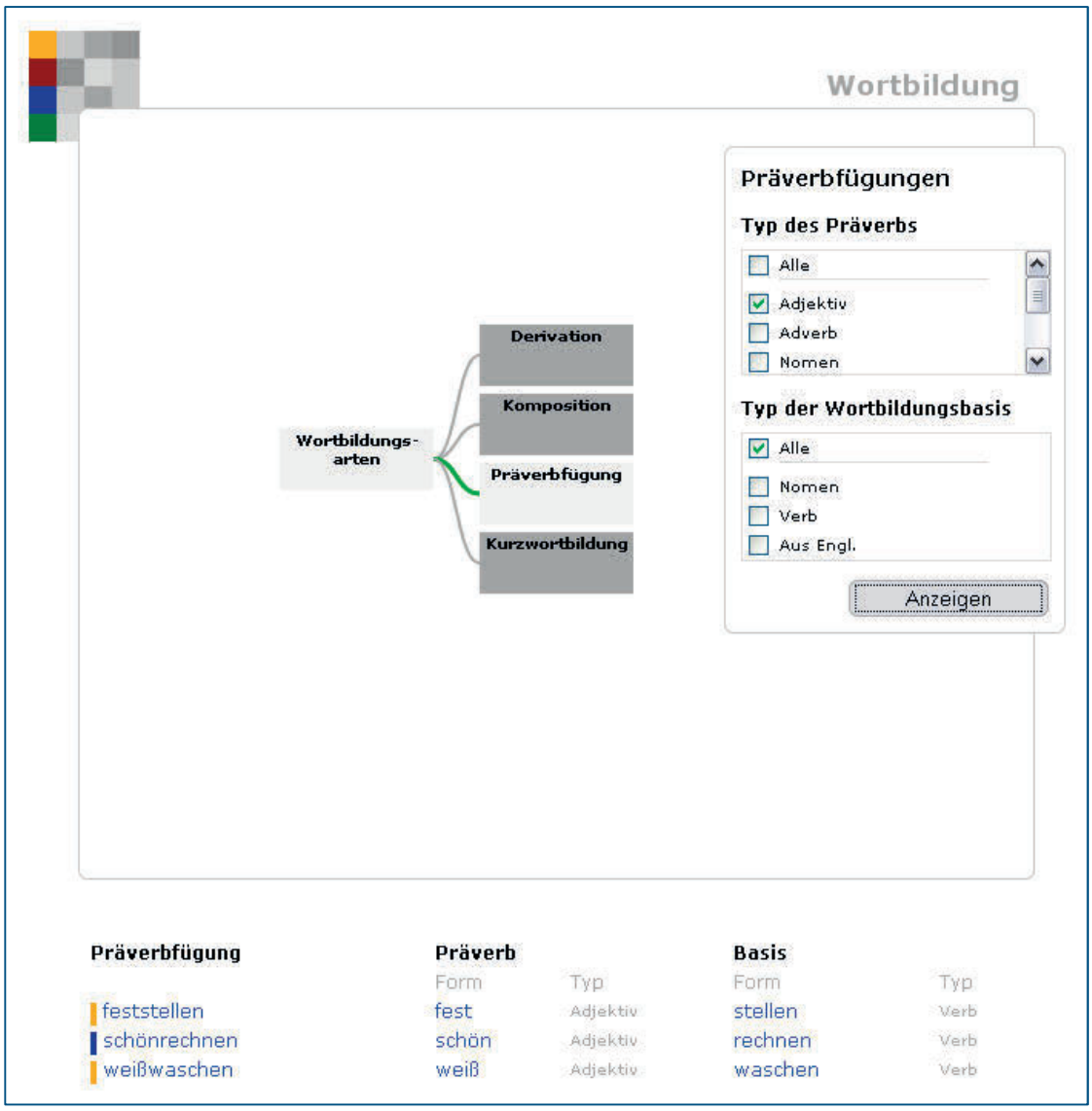


Abb. 7: Erweiterte OWID-Suche nach Präverbfügungen

Mit dieser neuartigen Form von erweiterten Suchen soll OWID als Portal für wissenschaftliche Lexikographie wie auch als Experimentierplattform dienen, in der neue Darstellungsmöglichkeiten erprobt und anhand empirischer Wörterbuchbenutzungsforschung (s. www.benutzungsforschung.de) auf ihre Anwendbarkeit überprüft werden können.

Abb. 8: Erweiterte OWID-Suche nach Kontaminationen

4. Unterstützung beim redaktionellen Arbeiten

Auch bei der redaktionellen Arbeit sollte die granulare, strenge Modellierung eine Unterstützung sein; so war der Anspruch bei der Entwicklung der DTDs. Ein wichtiger Bereich dabei ist die Erarbeitung, Verwaltung und Konsistenzsicherung der Verweisstrukturen lexikographischer Daten (vgl. Müller-Spitzer 2007b). Um zu verdeutlichen, wie in OWID, in diesem Fall genauer in *elexiko*, die Lexikographen dabei formal unterstützt werden, soll hier als ein Beispiel die Verwaltung der sinn- und sachverwandten Wörter demonstriert werden.

In *elexiko* werden zu allen Stichwörtern bzw. ihren Einzelbedeutungen (Lesarten) möglichst exhaustiv sinnverwandte Stichwörter (ggf. mit zugehörigen Einzelbedeutungen) korpusbasiert erarbeitet und im Wortartikel verzeichnet (vgl. u. a. Storjohann 2006 und 2011). Abbildung 9

zeigt als ein Beispiel einen Ausschnitt der sinnverwandten Wörter der Lesart ‚Material' im Wortartikel Holz.

Abb. 9: Ausschnitt aus dem *elexiko*-Wortartikel Holz, Lesart ‚Material', Bereich ‚Sinnverwandte Wörter'

Die Vernetzungsstruktur in *elexiko* ist demnach sehr umfangreich und ausführlich. Zur Verdeutlichung der Dimension: 1.250 ausgearbeitete Wortartikel enthalten 26.488 Relationspartner insgesamt, d.h., im Durchschnitt werden in *elexiko* 21 Relationspartner pro Stichwort verzeichnet. In den XML-Instanzen werden im Bereich der sinnverwandten Wörter die Typen der Vernetzung genau codiert, also ob es sich um Synonyme, inkompatible Partner oder komplementäre Partner etc. handelt, evtl. Kommentare zu Vernetzungen werden skopusgenau abgespeichert, und die ID des Zielstichworts bzw. der zugehörigen Lesart wird festgehalten. Letzteres ist besonders schwierig konsistent zu halten. *elexiko* ist ein im Aufbau befindliches Wörterbuch, d.h., Wortartikel werden kontinuierlich erarbeitet. Die Vernetzungen im Bereich der sinnverwandten Wörter sollen möglichst lesartenbezogen sein. Wenn nun aber beispielsweise der Artikel „Holz" erarbeitet und „Beton" als inkompatibler Partner verzeichnet wird, „Beton" aber noch nicht bearbeitet ist, dann kann auch keine Einzelbedeutung als Zieladresse angegeben werden. Wenn der Wortartikel „Beton" allerdings zu einem späteren Zeitpunkt bearbeitet wird, wünscht man sich als Lexikographin eine Information darüber, ob das Stichwort „Beton" bereits als Ziel in anderen Artikeln genannt wird. Genauso sollte überprüft werden können, ob Synonymverweise immer in beide Richtungen angelegt sind, wie dies konzeptionell gewünscht ist.

Die Modellierung kann für diese Anfragen die Basis liefern, indem alle relevanten Informationen in der XML-Struktur codiert sind. Allerdings ist zur Auswertung für die Lexikographen eine gesonderte Software nötig. Ein solcher Vernetzungsmanager wurde für *elexiko* im Rah-

men des Projekts BZV*elexiko* entwickelt.[31] Die Arbeit mit diesem Vernetzungsmanager kann grob folgendermaßen skizziert werden: Arbeitet ein Lexikograph im XML-Editor an einem Wortartikel (wie hier an „Holz“) und öffnet dann den Vernetzungsmanager, werden ihm alle eingehenden sowie alle aus dem Wortartikel ausgehenden Vernetzungen angezeigt. In der Spalte „Status“ wird zu jeder dieser Vernetzungen vermerkt, ob sie korrekt ist oder ob beispielsweise die Ziel-ID nicht korrekt ist oder die Vernetzung in der einen Richtung lesartenbezogen, in der anderen aber lesartenübergreifend ist etc. (vgl. Abbildung 10). Möchte der Lexikograph fehlerhafte Vernetzungen korrigieren, unterstützt der Vernetzungsmanager die Arbeit dahingehend, dass die relevanten Teile der Zielinstanz über den Manager geladen, korrigiert und wieder in die Datenbank eingecheckt werden können. Basis dafür, dass diese Abfragen trotz hoher Komplexität sehr performant sind, ist eine Linkdatenbank, die alle relevanten Extrakte aus den XML-Instanzen beinhaltet und über die die Abfragen des Vernetzungsmanagers laufen (vgl. Meyer/Müller-Spitzer 2010).

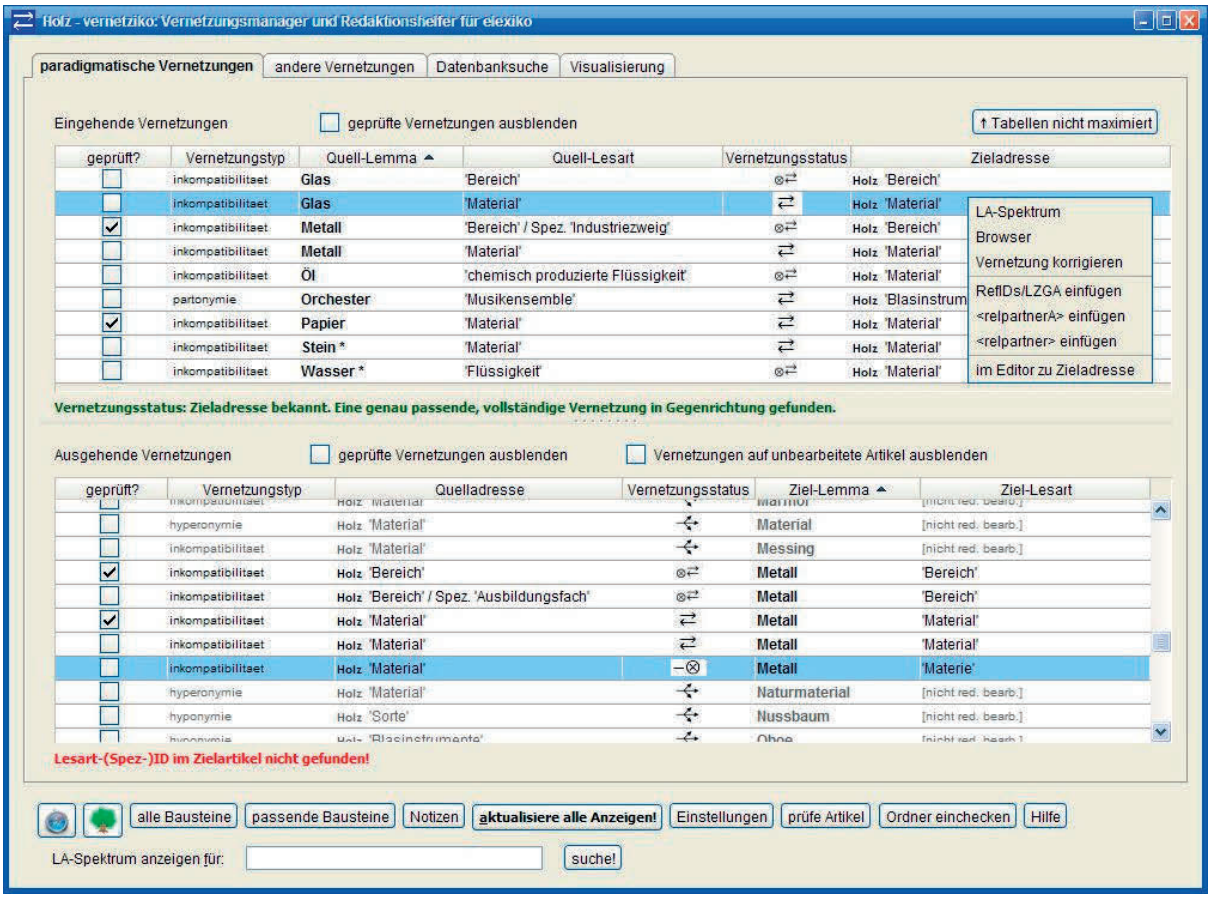


Abb. 10: Bildschirmansicht des Vernetzungsmanagers (Übersicht über ein- und ausgehende Vernetzungen)

Der hier dargestellte Vernetzungsmanager bietet darüber hinaus erweiterte Datenbanksuchen, die vor allem für die leitenden Lexikographen von Interesse sind. So können inhaltliche XPath-Abfragen mit weiteren Informationen aus der Datenbank wie Bearbeitungsstatus oder Bearbeitungszeit kombiniert werden, Gruppen von Instanzen können zusammen ein- und ausgecheckt werden oder Suchen mit selbstgewählten Extrakten aus den XML-Instanzen können spezifiziert werden (vgl. Abbildung 11). Seit der Einführung dieses Vernetzungsmanagers hat sich die redaktionelle Arbeit in *elexiko* daher deutlich vereinfacht und die Konsistenz der Da-

[31] Peter Meyer hat sowohl den Vernetzungsmanager als auch das Tool zur graphischen Visualisierung von Vernetzungen entwickelt.

ten hat sich erheblich – auch durch intensive redaktionelle Nacharbeiten – verbessert. Die Basis durch die genaue Modellierung war von Anfang an gegeben, allerdings fehlte bis vor kurzem eine entsprechende Softwareunterstützung.

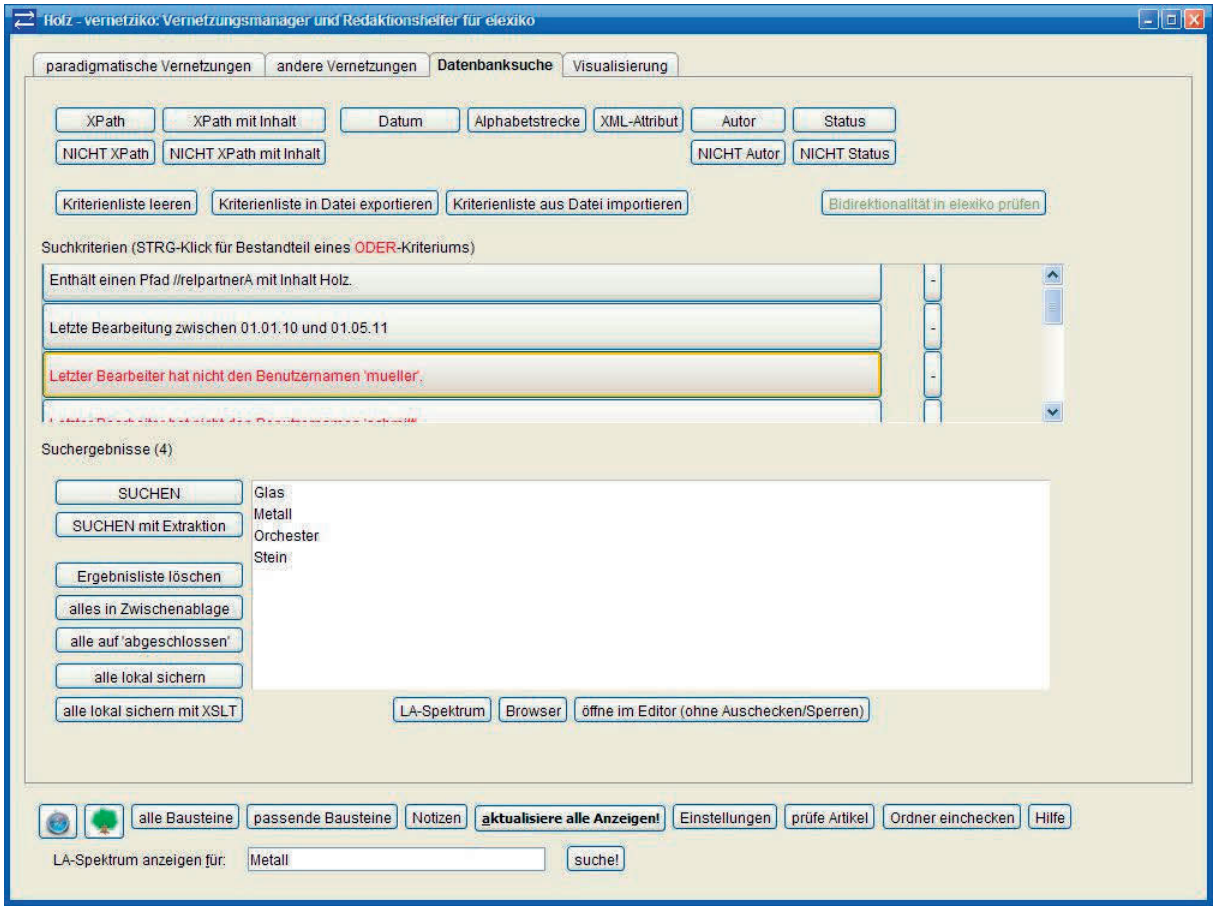


Abb. 11: Bildschirmansicht des Vernetzungsmanagers (Erweiterte Datenbanksuche)

Durch die Speicherung aller vernetzungsrelevanten Daten in einer separaten Linkdatenbank sind auch andere Darstellungen der Vernetzungen gut zu entwickeln und performant abfragbar. So haben wir beispielsweise eine graphische Visualisierung der sinnverwandten Wörter als Experimentierplattform für interne Zwecke (vgl. Abbildung 12) erarbeitet. Diese wird online noch keinem Benutzer zur Verfügung gestellt, weil noch nicht klar ist, für welche Benutzungssituation das Tool gewinnbringend einzusetzen ist.

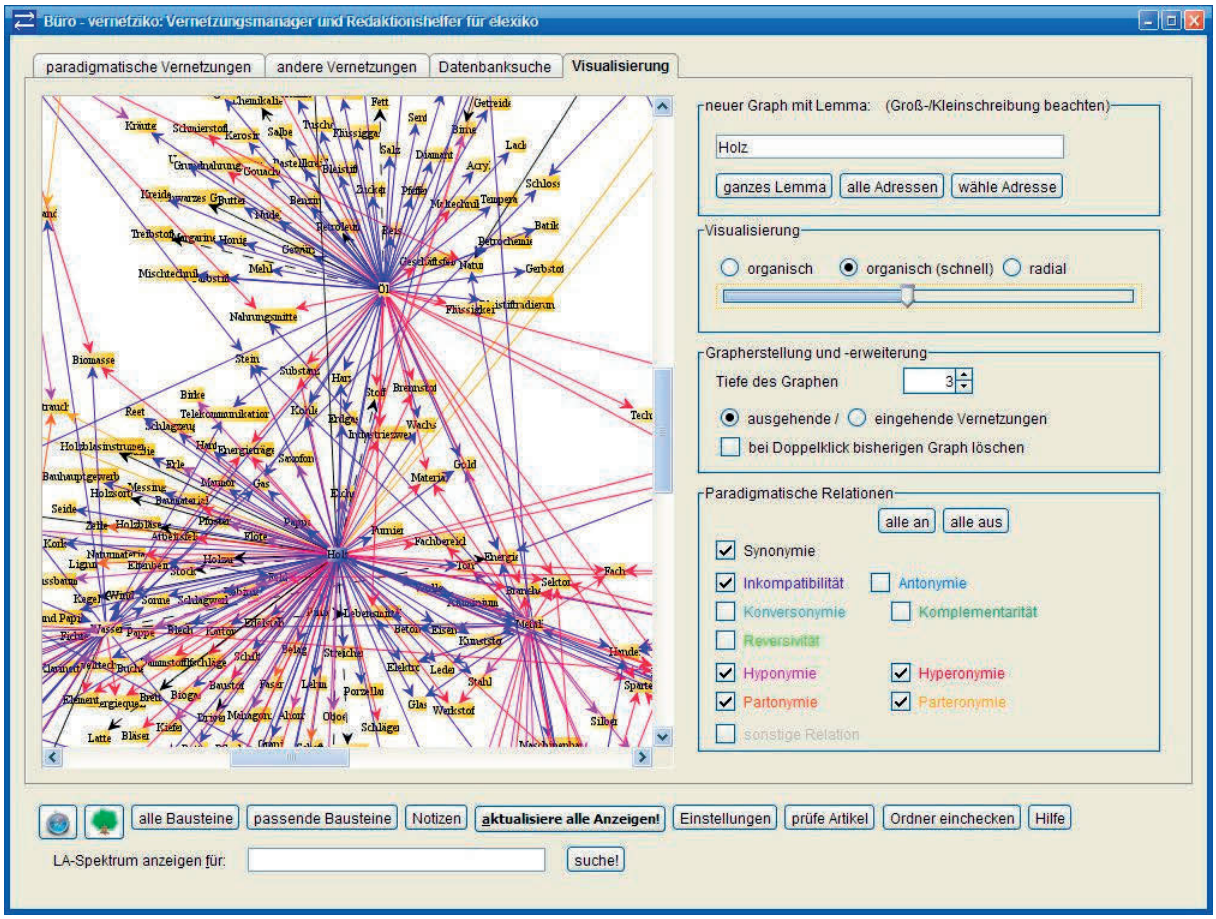


Abb. 12: Ausschnitt aus einer Bildschirmansicht des elexPlorer (Tool zur graphischen Visualisierung der sinnverwandten Wörter)

5. Vor- und Nachteile des Modellierungsansatzes

Nachdem in den vorangegangenen Abschnitten die Arbeit mit der Modellierung in OWID verdeutlicht wurde und damit auch mehr die Vorteile des gewählten Modellierungsansatzes im Vordergrund standen, sollen abschließend kurz Vor-, aber auch Nachteile des hier vorgestellten Ansatzes gegenübergestellt werden. Dabei werden zwei wesentliche Aspekte herausgegriffen: die maßgeschneiderte Modellierung und die Ausrichtung als Inhaltsstrukturmodellierung.

5.1 Maßgeschneiderte Modellierung

Eine maßgeschneiderte Modellierung zu entwickeln ist in einer Zeit zunehmender Standardisierungsbemühungen und Gründung von Infrastrukturprojekten mit dem Ziel, ein möglichst hohes Maß an Austauschbarkeit von Daten zu gewährleisten (wie TextGrid, CLARIN, D-Spin etc.[32]), in gewissem Sinne unmodern. Trotzdem haben wir mit dieser Entscheidung nur gute Erfahrungen gemacht. Eine maßgeschneiderte Modellierung ermöglicht einen passgenauen Zuschnitt auf die Erfordernisse der einzelnen Wörterbücher in einem Portal. Gerade wenn die lexikographischen Daten kontinuierlich erarbeitet werden, ist dies ein enormer Vorteil. Wollte

[32] Siehe www.textgrid.de, www.clarin.eu, http://weblicht.sfs.uni-tuebingen.de/.

man eine Standard-Modellierung so genau auf die individuellen Bedürfnisse der einzelnen lexikographischen Ressourcen anpassen, wäre auch eine ursprünglich standardbasierte Modellierung von einer maßgeschneiderten nicht weit entfernt. Sehr anders sieht die Situation bei der Strukturierung retrodigitalisierter Wörterbücher aus (vgl. den Beitrag von Hildenbrandt in diesem Band). Allerdings muss man zu allen Standard-Modellierungen wie der TEI (www.tei.org) bemerken: Der Vorteil einer Standard-Modellierung soll die Austauschbarkeit von Daten sein; gleichzeitig muss eine Standard-Modellierung hinreichend offen sein, um von unterschiedlichsten Projekten angewendet werden zu können. Diese beiden Pole stehen oft im Widerstreit zueinander. Man kann davon ausgehen, dass zwei lexikographische Projekte, die unabhängig voneinander beispielsweise die P5-Richtlinien der TEI anwenden, ihre Daten trotzdem nicht ohne Weiteres austauschen können, da diese Richtlinien eben sehr unterschiedlich angewendet werden können. Trotzdem sollte man sich immer über die Standards informieren (so wie es am IDS auch getan wurde), da meist viele Fachleute an der Entwicklung beteiligt sind und man diese Richtlinie als Orientierung benutzen kann, auch wenn die gesamte Modellierung maßgeschneidert ist.

Da die für OWID gewählte Modellierung allerdings möglichst feingranular und genau ist, lässt sich diese maßgeschneiderte Modellierung jederzeit in eine standardbasierte Modellierung z.B. analog zu den TEI-Richtlinien überführen, da diese Standard-Modellierungen immer sehr viel allgemeiner gehalten sind. Für die Beteiligung an Infrastrukturprojekten wurde eine solche Migration bereits in der Praxis erprobt.

5.2 Inhaltsstrukturmodellierung

Der Ansatz einer Inhaltsstrukturmodellierung, d.h. die möglichst granulare Strukturierung aller lexikographischen Angaben, strikt orientiert am inhaltlichen Gehalt der Daten, hat wie oben ausgeführt die Vorteile, die Lexikographen bei der Einhaltung der formalen Artikelstruktur bestmöglich zu unterstützen sowie die Basis für sehr flexible Zugriffsmöglichkeiten für Lexikographen und Endbenutzer zu legen. Dieser Ansatz birgt allerdings auch Nachteile: Der Aufwand, aus solchen granular gegliederten Daten, die von ihrem Aufbau her mehr an den linguistischen Strukturen als an der Gliederung des Wortartikels im Online-Wörterbuch orientiert sind, eine Präsentation (z.B. über XSLT-Stylesheets) zu entwickeln, ist sehr hoch. Beispielsweise muss für jedes der 400 Elemente und zugehörigen Attribute von *elexiko* festgelegt werden, wie die entsprechende Information in einer Online-Ansicht dargestellt werden soll. Außerdem ist die Modellierung rein am Inhalt orientiert, d.h., viele Elemente müssen in eine andere Reihenfolge etc. gebracht werden, um im Wortartikel in der gewünschten Form zu erscheinen. Diese grundlegende Schwierigkeit wird für *elexiko* verschärft dadurch, dass zunächst die Modellierung der Daten entwickelt wurde (an der inhaltlichen Struktur orientiert) und erst danach die Gliederung der Wortartikel für die Online-Darstellung festgelegt wurde. Die Schere zwischen der Gliederung der Daten in der Datenbasis und der Online-Ansicht klafft daher an manchen Stellen weit auseinander.

Ein weiterer, etwas diffiziler zu erklärender Nachteil kommt hinzu: Der Anspruch der Inhaltsstrukturmodellierung ist es, Inhalte zu modellieren und die Skopusbeziehungen beispielsweise von Angaben und dazugehörigen Kommentaren so genau wie möglich abzubilden, losgelöst von Aspekten der Präsentation. Die granulare Strukturierung erlaubt es, aus einer so strukturierten Datenbasis für eine Präsentation einen lexikographischen Text ‚zusammenzubauen'. Wenn nun aber ein Wörterbuch wie *elexiko* oder das Neologismenwörterbuch über Jahre erarbeitet wird und die Wortartikel in der Online-Ansicht eine fest definierte Gestalt haben, kann über die Jahre Folgendes passieren: Die Lexikographen denken bei der Datenerarbeitung

und -strukturierung mehr daran, wie die Daten im Wortartikel online aussehen sollen, als daran, wo sie eigentlich inhaltlich hingehören. Ein Beispiel: Zu einem Relationspartner aus dem Bereich der sinnverwandten Wörter soll ein Kommentar gegeben werden. Fiktiv sei es so, dass online sowohl ein Kommentar zu allen Relationspartnern dieses Typs wie auch die Kommentare zu einzelnen Partnern an der derselben Stelle erscheinen können. In einem solchen Fall kann es passieren, dass ein Lexikograph den Kommentar an der falschen Stelle in der Artikelstruktur eingibt (also an der Stelle, an der eigentlich nur ein Kommentar zu allen Relationspartnern stehen darf), weil er nur überprüft, ob der Kommentar online korrekt erscheint. Zwar wird auch das XML-Tagging der Artikel in *elexiko* Korrektur gelesen, aber gerade diese Skopusbeziehungen sind sehr schlecht zu überblicken. Dies ist nur ein subtiles Beispiel für dieses Phänomen, aber gerade das Denken von der Präsentation her kann zu einem Problem bei der Inhaltsstrukturmodellierung werden, ist aber bei einer langjährigen lexikographischen Routine kaum zu vermeiden.

6. Schlussbemerkung

Für das Wörterbuchnetz OWID hat sich der hier dargestellte Modellierungsansatz bewährt. Die Modellierung ist ausgerichtet auf unterschiedliche lexikographische Ressourcen, die neu erarbeitet werden, die bestimmte Angabebereiche teilen und so zum Teil gemeinsame Inhaltsstrukturen haben, zum Teil individuell verschieden strukturiert sein müssen. Der vorliegende Beitrag sollte einen Einblick in die praktische Arbeit von OWID bieten und so möglichst die Diskussion zwischen lexikographischen Projekten mit unterschiedlichen Modellierungsansätzen befruchten.

7. Literatur

Benutzungsforschung. Internet: www.benutzungsforschung.de. (Stand: Oktober 2011).

CLARIN – Common Language Resources and Technology Infrastructure. Internet: http://www.clarin.eu. (Stand: Oktober 2011).

elexiko (2003ff.), in: OWID – Online Wortschatz-Informationssystem Deutsch, hg. v. Institut für Deutsche Sprache, Mannheim. Internet: www.owid.de/elexiko_/index.html. (Stand: Oktober 2011).

Engelberg, Stefan/Müller-Spitzer, Carolin (im Erscheinen): Dictionary portal. In: Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography, hg. von Gouws, Rufus H./Heid, Ulrich/Schweickard, Wolfgang/Wiegand, Herbert Ernst. Berlin/New York.

E-ValBU – Das elektronische Valenzwörterbuch deutscher Verben. Internet: http://hypermedia2.ids-mannheim.de/evalbu/index.html. (Stand: Oktober 2011).

Feste Wortverbindungen (2007ff.), in: OWID – Online Wortschatz-Informationssystem Deutsch, hg. v. Institut für Deutsche Sprache, Mannheim. Internet: www.owid.de/Wortverbindungen/index.html. (Stand: Oktober 2011).

Harras, Gisela/Winkler, Edeltraud/Erb, Sabine/Proost, Kristel (2004): Handbuch deutscher Kommunikationsverben. Teil 1: Wörterbuch. (= Schriften des Instituts für Deutsche Sprache 10.1). Berlin/New York.

Herberg, Dieter/Steffens, Doris/Tellenbach, Elke (1997): Schlüsselwörter der Wendezeit. Wörter-Buch zum öffentlichen Sprachgebrauch 1989/90. (= Schriften des Instituts für deutsche Sprache 6). Berlin/New York.

IDS – Institut für Deutsche Sprache. Internet: http://www.ids-mannheim.de. (Stand: Oktober 2011).

Klosa, Annette (Hg.) (2008): Lexikografische Portale im Internet. (= OPAL – Online publizierte Arbeiten zur Linguistik 1/2008). Mannheim.

Klosa, Annette (2011): Korpusgestützte Angaben zu Grammatik und Wortbildung. In: Klosa, Annette (Hg.): *elexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. (= Studien zur deutschen Sprache 55). Tübingen, S. 145-156.

Meyer, Peter/Müller-Spitzer, Carolin (2010): Consistency of Sense Relations in a Lexicographic Context. In: Barbu Mititelu, Verginica/Pekar, Viktor/Barbu, Eduard (Hg.): Proceedings of the Workshop „Semantic Relations. Theory and Applications“, 18 May 2010, at the International Conference on Language Resources and

Evaluation (LREC) 2010, Malta. Internet: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W9.pdf. (Stand: Oktober 2011).

Müller-Spitzer, Carolin (2007a): Der lexikografische Prozess. Konzeption für die Modellierung der Datenbasis. (= Studien zur deutschen Sprache 42). Tübingen.

Müller-Spitzer, Carolin (2007b): Vernetzungsstrukturen lexikografischer Daten und ihre XML-basierte Modellierung. In: Hermes 38, S. 137-171.

Müller-Spitzer, Carolin (2008a): Der texttechnologische Aufbau von OWID. In: Klosa, Annette (Hg.): Lexikografische Portale im Internet. (= OPAL – Online publizierte Arbeiten zur Linguistik 1/2008). Mannheim, S. 45-55.

Müller-Spitzer, Carolin (2008b): The Lexicographic Portal of the IDS. Connecting Heterogeneous Lexicographic Resources by a Consistent Concept of Data Modelling. In: Proceedings of the 13th EURALEX International Congress. Euralex 2008. Barcelona, Spain (CD-ROM).

Müller-Spitzer, Carolin (2011): Der Einsatz einer maßgeschneiderten, feingranularen XML-Modellierung im lexikografischen Prozess. In: Klosa, Annette (Hg.): *elexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. (= Studien zur deutschen Sprache 55). Tübingen, S. 173-191.

OWID – Wortschatzinformationssystem Deutsch (2008ff.), hg. v. Institut für Deutsche Sprache, Mannheim. Internet: http://www.owid.de. (Stand: Oktober 2011).

Diskurswörterbuch 1945-55 (2007), in: OWID – Online Wortschatz-Informationssystem Deutsch, hg. v. Institut für Deutsche Sprache, Mannheim, www.owid.de/Diskurs1945-55/index.html. (Stand: Oktober 2011).

SprichWort-Plattform (2008-2010). Internet: http://www.sprichwort-plattform.org/sp/Sprichwort. (Stand: Oktober 2011).

Storjohann, Petra (2006): Sinnrelationen in Wörterbüchern – Neue Ansätze und Perspektiven. In: EliSe 2/2005, S. 35-61. Internet: http://www.uni-due.de/imperia/md/content/elise/ausgabe_2_2005_storjohann.pdf). (Stand: Oktober 2011).

Storjohann, Petra (2011): Paradigmatische Konstruktionen in Theorie, lexikografischer Praxis und im Korpus. In: Klosa, Annette (Hg.): *elexiko*. Erfahrungsberichte aus der lexikographischen Praxis eines Internetwörterbuchs. (= Studien zur deutschen Sprache 55). Tübingen, S. 99-129.

TEI – Text Encoding Initiative. Internet: http://www.tei-c.org/index.xml. (Stand: Oktober 2011).

TextGrid – Vernetzte Forschungsumgebung in den eHumanities. Internet: http://www.textgrid.de. (Stand: Oktober 2011).

WebLicht – Deutsche Sprachressourcen-Infrastruktur D-SPIN. Internet: http://weblicht.sfs.uni-tuebingen.de/. (Stand: Oktober 2011).

Wortverbindungen online. Plattform des Projekts Usuelle Wortverbindungen. Internet: http://wvonline.ids-mannheim.de/. (Stand: Oktober 2011).

Datenmodelle und Datenformate für die Modellierung des Fußballwortschatzes im Kicktionary

Thomas Schmidt thomas.schmidt@ids-mannheim.de, Tel.: +49 621 1581-304

1. Einleitung

Das Kicktionary ist ein dreisprachiges (deutsch-englisch-französisches) elektronisches Wörterbuch der Fußballsprache. Es basiert auf einem Korpus von geschriebenen Fußballberichten und (in geringerem Umfang) gesprochenen Fußballkommentaren und nutzt die Ideen der Framesemantik (Fillmore 1982, Fillmore et al. 2003) sowie der lexikalischen Relationen (Fellbaum 1998) zur Strukturierung des Wortschatzes. Verschiedene Aspekte der Erstellung, Präsentation und Nutzung des Kicktionary sind in Schmidt (2008, 2009 und 2010) dargestellt. Im vorliegenden Beitrag konzentriere ich mich auf die Frage, welche Datenmodelle und welche Datenformate zur Modellierung des Wortschatzes im Kicktionary zum Einsatz kamen. Zu diesem Zweck möchte ich einleitend zunächst mein Verständnis dieser drei Begriffe – Datenmodell, Datenformat und Modellierung – näher erläutern.

Unter dem *Datenmodell* des Kicktionary verstehe ich die Struktur, mittels derer die Gesamtheit der Wörterbucheinheiten und deren Beziehungen untereinander erfasst werden. Das Datenmodell ist abstrakt, d.h., es lässt sich zwar (symbolisch/verbal) beschreiben, hat aber selbst keine konkrete physikalische Manifestation, sondern wird nur beispielsweise in Form des zugehörigen Datenformats oder der nutzerorientierten Visualisierung des Wörterbuchs direkt erfassbar. Das *Datenformat* ist in diesem Sinne also die (oder eine) konkrete physikalische Repräsentation des Datenmodells. Es hat typischerweise die Form einer oder mehrerer XML-Dateien oder einer relationalen Datenbank. Mit dieser Unterscheidung folge ich der in der Datenbanktechnologie üblichen Differenzierung zwischen physikalischer und logischer Ebene der Datenverarbeitung, der als drittes die Anwendungsebene zur Seite gestellt wird (vgl. z.B. Date 1995). Unter *Modellierung* schließlich verstehe ich den empirischen Prozess, in dem das Modell aufgebaut, ggf. modifiziert und mit Inhalt gefüllt wird. Konkreter: Unter den Begriff der Modellierung fasse ich all diejenigen Vorgänge in der Wörterbucherstellung, in denen Wörterbucheinträge erstellt, ergänzt und geändert und in die (sich eventuell ebenfalls ändernde) Struktur des Wörterbuchs eingeordnet werden. Das Datenmodell, das in einem Datenformat physikalisch repräsentiert ist, ist in diesem Sinne das *Produkt*, die Modellierung der *Prozess* der Wörterbucherstellung.[33]

Im folgenden Abschnitt beschreibe ich also zunächst Datenmodell und -format(e) des Kicktionary. Anschließend stelle ich den Modellierungsprozess und einige damit zusammenhängende Schwierigkeiten dar. Schließlich skizziere ich zwei Änderungen am Modell bzw. am Modellierungsprozess, die mir geeignet scheinen, zumindest einen Teil dieser Schwierigkeiten zu überwinden.

[33] Ich wurde zu Recht darauf hingewiesen, dass dieses Verständnis des Begriffs „Modellierung“ weiter gefasst ist als in der Metalexikographie üblich – vgl. z.B. Geeb (2001), S. 29: „Datenmodellierung sei verstanden als der Entwurf von (lexikographischen) Dateneinheiten und ihren funktionalen Beziehungen ohne Rücksichtnahme auf bereits bestehendes Datenmaterial.“ Wie ich weiter unten darlege, ist es aber gerade die Wechselwirkung vorgegebener Datenstrukturen mit ebenjenem „bereits bestehenden [oder neu hinzukommendem] Datenmaterial“, die sich bei der Erstellung des Kicktionary als interessantes Problem erwiesen hat. Vor diesem Hintergrund erscheint mir die Erweiterung der Bedeutung des Begriffs „Modellierung“ für die Zwecke dieses Beitrags zu rechtfertigen.

2. Datenmodell und Datenformate

Zentrale Einheit im Datenmodell des Kicktionary ist – wie im Berkeley FrameNet – die LEXIKALISCHE EINHEIT (abgekürzt: LU), die Ruppenhofer et al. (2010, S. 5) als „a pairing of a word form with a meaning“ definieren.[34]Die lexikalische Einheit ist insofern zentral, als sich jeder andere Modellbestandteil zu ihr in eine direkte Beziehung setzen lässt – jede andere Wörterbucheinheit ist entweder Bestandteil einer lexikalischen Einheit oder dient dazu, lexikalische Einheiten zueinander in Beziehung zu setzen oder in größere Struktureinheiten zusammenzufassen.

Zu einer lexikalischen Einheit gehören erstens (obligatorisch) Angaben zur Sprache und zum PART-OF-SPEECH, zweitens (optional) eine DEFINITION und drittens (obligatorisch) eine Reihe von KORPUSBEISPIELEN, in denen (u.a.) die verwendete Form der lexikalischen Einheit markiert ist.

Lexikalische Einheit:	*tunneln*
Sprache:	Deutsch
Part-Of-Speech:	Verb
Definition:	Der ballführende Spieler behauptet den Ballbesitz gegen einen angreifenden Gegenspieler, indem er ihm den Ball zwischen den Beinen hindurch spielt.
Korpusbeispiele:	
(1)	Diogo Rincón tunnelte Paul Freier im Strafraum und sein Schuss trudelte […] an Jörg Butt vorbei und landete in Netz.
(2)	[…] nur wenige Sekunden später war es Duff, der nach einer herrlichen Kombination zwischen Kezman und Cole […] den spanischen Keeper zum 3:0 **tunnelte**.
(3)	Ailton tunnelte Chris an der Strafraumgrenze und spielte so Klasnic frei.

Abb. 1: Beispiel für eine lexikalische Einheit mit Korpusbeispielen, vgl. http://www.kicktionary.de/LUs/Beat/LU_391.html

Jede lexikalische Einheit ist genau einem FRAME zugeordnet, der seinerseits Bestandteil genau einer SCENE ist. Eine Scene wird dabei verstanden als das (u.U. nonverbale) Wissen über einen prototypischen Handlungsablauf; ein Frame fasst alle lexikalischen Einheiten zusammen, mittels derer ein bestimmter Aspekt dieses Handlungsablaufes aus einer bestimmten Perspektive sprachlich dargestellt werden kann. Zu einer Scene gehört eine Charakterisierung des Handlungsablaufes selbst sowie der typischerweise an ihr beteiligten Aktanten und Gegenstände. Letztere werden FRAME-ELEMENTE genannt.

[34] Eine Form korrespondiert hier also immer nur mit einer Einzelbedeutung. Eine Einheit wie das „Stichwort“ in einem klassischen Wörterbuch, unter der mehrere Einzelbedeutungen zusammengefasst werden, gibt es beim Kicktionary nicht als explizite Einheit im Modell.

Scene:	One-on-One
Handlungsablauf:	Der ballführende Spieler wird von einem Gegenspieler, der versucht, in Ballbesitz zu gelangen, angegriffen. Ein Zweikampf endet, indem entweder der ballführende Spieler den Ballbesitz behauptet oder der angreifende Spieler in Ballbesitz gelangt.
Frame-Elemente:	BALLFÜHRENDER_SPIELER, ANGREIFENDER_SPIELER, BALL
Zugehörige Frames:	Challenge (LUs: *angreifen*, *attackieren*, *bedrängen*, *stören* etc.) Beat (LUs: *ausspielen*, *umdribbeln*, *vernaschen*, *tunneln* etc.) Deny (LUs: *abgrätschen*, *abjagen*, *blocken*, *stoppen* etc.)

Abb. 2: Beispiel für eine Scenemit zugehörigen Frames, vgl. http://www.kicktionary.de/One_On_One_Scenario.html

Die Frame-Elemente finden sich in den Korpusbeispielen als Annotationen der Mitspieler der lexikalischen Einheit wieder.

(1)	[Diogo Rincón]$_{\text{BALLFÜHRENDER SPIELER}}$ tunnelte [Paul Freier]$_{\text{ANGREIFENDER SPIELER}}$ im Strafraum und sein Schuss trudelte […] an Jörg Butt vorbei und landete in Netz.
(2)	[…] nur wenige Sekunden später war es [Duff]$_{\text{BALLFÜHRENDER SPIELER}}$, der nach einer herrlichen Kombination zwischen Kezman und Cole […] [den spanischen Keeper]$_{\text{ANGREIFENDER SPIELER}}$ zum 3:0 **tunnelte**.
(3)	[Ailton]$_{\text{BALLFÜHRENDER SPIELER}}$ tunnelte [Chris]$_{\text{ANGREIFENDER SPIELER}}$ an der Strafraumgrenze und spielte so Klasnic frei.

Abb. 3: Korpusbeispiele mit annotierten Frame-Elementen

Die einzelnen Wörterbucheinträge – lexikalische Einheiten mit den zugehörigen Spezifizierungen und Beispielen – werden also einerseits in einer Hierarchie aus Scenes und Frames organisiert.

Andererseits werden, dem Ansatz von WordNet folgend, bedeutungsgleiche lexikalische Einheiten zu SYNSETS zusammengefasst, wobei das Kicktionary als mehrsprachiges Wörterbuch auch die Übersetzungsäquivalenz als eine Form der Synonymie behandelt. Zwischen Synsets können darüber hinaus die SEMANTISCHEN RELATIONEN Hyperonymie/Hyperonymie (Ober- bzw. Unterbegriff bei Substantiven), Meronymie/Holonymie (Teil-Ganzes-Beziehung bei Substantiven) sowie Troponymie (Analogon zur Hyperonymie bei Verben) im Datenmodell festgehalten werden.

Synset:	{*tunneln / to nutmeg*}
Troponyme:	{*ausspielen*, *sich durchsetzen / beat*, *round / éliminer*}
Synset:	{*Innenverteidiger / central defender*, *centre-back / défenseur central*}
Hyperonyme:	{*Verteidiger*, *Abwehrspieler / defender / arrière*, *défenseur*}
Holonyme:	{*Verteidigung*, *Abwehr / defence*, *backline / arrièregarde*, *défense*}

Abb. 4: Beispiele für Synsets und semantische Relationen

Durch Verkettung der transitiven semantischen Relationen entstehen weitere hierarchische Organisationen des Wortschatzes, die im Kicktionary KONZEPTHIERARCHIEN genannt werden.

Mannschaft, Team / side, squad / équipe, formation
Schlussmann, Torhüter / custodian, goalkeeper / gardien, portier
Abwehr, Verteidigung / backline, defence / arrière-garde, défense
Innenverteidigung / central defence / défense centrale
Innenverteidiger / central defender / défenseur central
Mittelfeld / midfield / milieu de terrain
Mittelfeldspieler / midfielder / milieu
Angriff, Sturm / attack / attaque
Mittelstürmer / centre-forward / avant-centre

Abb. 5: Beispiel für eine Konzepthierarchie, basierend auf der semantischen Relation Meronymie/Holonymie, vgl. http://www.kicktionary.de/CONCEPT_HIERARCHIES/Hypernymy_Individual_Actors.html

Anders als bei der Scenes-Frames-Hierarchie gilt nicht, dass eine lexikalische Einheit genau einer Konzepthierarchie zugeordnet sein muss. Es tritt sowohl der Fall auf, dass eine lexikalische Einheit gar nicht in einer solchen Hierarchie auftaucht (weil es keine weiteren lexikalischen Einheiten gibt, mit denen sie in einer semantischen Relation steht), als auch der Fall, dass sie Bestandteil mehrerer solcher Hierarchien ist (weil sie z.B. sowohl Holonym als auch Hyperonym von anderen lexikalischen Einheiten ist).

Abbildung 6 stellt noch einmal die wichtigsten Einheiten im Kicktionary-Datenmodell und ihre Beziehungen zueinander dar.

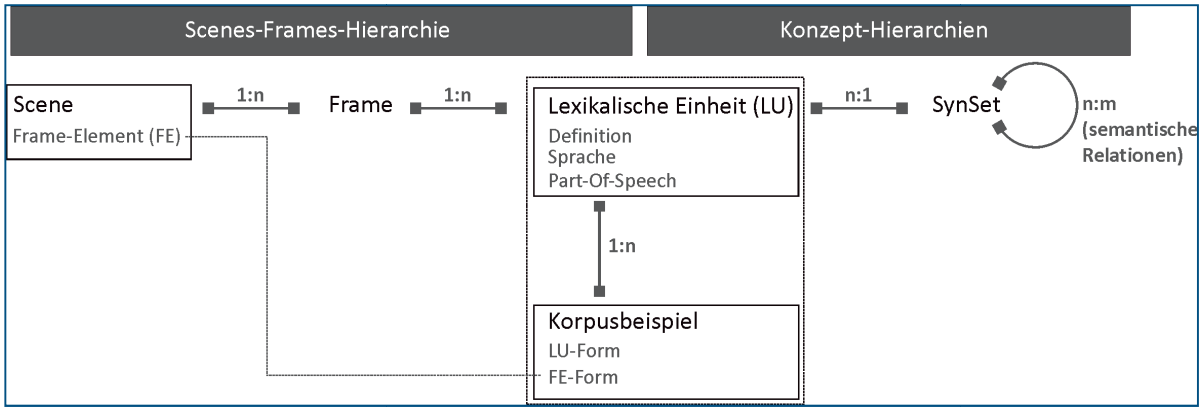


Abb. 6: Schema des Kicktionary-Datenmodells

Auf der physikalischen Ebene wurde dieses Modell während der Erstellung in Form einer einfachen XML-Datei repräsentiert, wie in Abbildung 7 illustriert. Da das Kicktionary von vornherein als ein zeitlich befristetes Projekt mit einem einzigen Bearbeiter angelegt war, werden nicht alle strukturellen Beschränkungen des Modells auf der Ebene des Dateiformats abgebildet, d.h., die zugehörige DTD stellt zwar z.B. sicher, dass ID-Bezüge in den Frame-Element-Annotationen (Attribute `@fe-idref`) auf eine existierende ID (Attribute `@name` in den Elementen `<argument>`) verweisen, nicht aber, dass dieser Bezug sich auch innerhalb der betreffenden lexikalischen Einheit befindet oder über verschiedene LUs innerhalb eines Frames konsistent benannt ist. Die Integrität des Datenformats im Bezug auf das Modell wurde stattdessen teilweise durch eine entsprechende Funktionalität in der Bearbeitungssoftware (ein eigens für das Projekt entwickelter Editor mit Konkordanzfunktion für das Korpus, siehe Schmidt 2008), teilweise einfach durch regelmäßige manuelle Kontrollen sichergestellt.

```
<LU_REPORT id="LU_391">
<LEXICAL-UNIT scenario="One_On_One" frame="Beat"
        lu-id="tunneln.v" lang="de" wordclass="v" synset="TO_NUTMEG">tunneln</LEXICAL-
UNIT>
<arguments>
<argument name="PLAYER_WITH_BALL"/>
<argument name="AREA"/>
<argument name="OPPONENT_PLAYER"/>
</arguments>
<DEFINITION>Der ballführende Spieler (PLAYER) überwindet einen angreifenden Gegenspieler
        (OPPONENT_PLAYER), indem er ihm den Ball zwischen den Beinen hindurch spielt. Der
        Ort,an dem dieser Zweikampf stattfindet (LOCATION) kann erwähnt werden.
</DEFINITION>
<EXAMPLE lang="de">
<FE_REF fe-idref="PLAYER_WITH_BALL">Diogo Rincón</FE_REF>
<LU_REF lu-idref="tunneln.v">tunnelte</LU_REF>
<FE_REF fe-idref="OPPONENT_PLAYER">Paul Freier</FE_REF>
<FE_REF fe-idref="AREA">im Strafraum</FE_REF>
        und sein Schuss trudelte, begünstigt durch
        einen Platzfehler, an Jörg Butt vorbei und landete in Netz.
<EXAMPLE lang="de" source-element-id="p5" source-text-id="K_175c">
<FE_REF fe-idref="PLAYER_WITH_BALL">Ailton</FE_REF>
<LU_REF lu-idref="tunneln.v">tunnelte</LU_REF>
<FE_REF fe-idref="OPPONENT_PLAYER">Chris</FE_REF>
<FE_REF fe-idref="AREA">an der Strafraumgrenze</FE_REF> und spielte so Klasnic frei.
</EXAMPLE>
</LU_REPORT>
```

Abb. 7: XML-Repräsentation einer lexikalischen Einheit

3. Modellierung

Ausgangspunkt für die Modellierung des Fußballwortschatzes im Kicktionary waren einfache Wortlisten, die aus den Korpora der drei beteiligten Sprachen automatisch generiert und nach Häufigkeit der Vorkommen geordnet wurden. Auf der Grundlage dieser Wortlisten wurde dann für jede Sprache eine Liste von je ca. 100 lexikalischen Einheiten erstellt, die entweder spezifisch für die Fußballsprache sind oder zumindest in Fußballberichten in deutlich erhöhter Frequenz auftreten. Dieses Material bildete dann die Basis für die weitergehenden Analysen.

Das weitere Vorgehen zum Aufbau der Scenes-Frames-Hierarchie bestand in einem iterativen Prozess: Das vorhandene Material lexikalischer Einheiten wurde zunächst anhand von Vorkommen im Korpus analysiert und nach Ähnlichkeiten im Bezug auf Argumentstrukturen und semantische Eigenschaften geordnet. Daraufhin wurden erste Scenes und Frames mit den zugehörigen Frame-Elementen definiert, für jede lexikalische Einheit eine Zuordnung zu einem Frame vorgenommen und Korpusbeispiele gemäß den betreffenden Frame-Elementen annotiert.

Meine Grundannahme war dabei, dass es nicht darum geht, *die* „korrekten“ Scenes und Frames der Fußballsprache zu ermitteln und zu beschreiben. Vielmehr gehe ich davon aus, dass es sich bei der Scenes- und Frames-Analyse um ein wissenschaftliches Modell handelt, das nicht an sich falsch oder richtig sein kann. Sein Wert bemisst sich eher danach, wie nützlich und handhabbar seine verkürzende Abbildung der Realität im Bezug auf einen bestimmten Erkenntnis- oder Anwendungszweck ist (vgl. dazu die Ausführungen zum pragmatischen Merkmal wissenschaftlicher Modelle in Stachowiak 1973). In diesem Sinne begreife ich Scenes und Frames in erster Linie als Mittel zur Makrostrukturierung eines Wörterbuchs (und nicht z.B. als Ansatz zur Beschreibung oder Erklärung kognitiver Strukturen oder als Basis für eine maschinelle Sprachverarbeitung). Ihr Sinn besteht für mich darin, die potentiell unüberschaubare Gesamtheit lexikalischer Einheiten dergestalt in kleinere Einheiten zu strukturieren, dass es dem Benutzer ermöglicht bzw. vereinfacht wird, Beziehungen im Wortschatz

zu erkennen und für die Sprachrezeption oder -produktion fruchtbar zu machen. Nützliche Frames sind demnach solche, die

- erstens eine hilfreiche „Portionierung“ des Gesamtwortschatzes leisten. Frames, die zu viele lexikalische Einheiten enthalten, sind nach diesem Kriterium ebenso defizitär wie Frames, die zu wenige lexikalische Einheiten enthalten. Es sei zur Klarstellung angemerkt, dass die Anzahl der Mitglieder eines Frames vor allem mit seinem Abstraktionsgrad oder dem der übergeordneten Scene variiert, die Framesemantik aber keine Aussagen über einen angemessenen Abstraktionsgrad macht.
- zweitens nur solche lexikalischen Einheiten enthalten, die bezüglich ihrer Semantik und Argumentstrukturen ausreichend homogen sind. Eine zu geringe Homogenität äußert sich beispielsweise darin, dass die annotierten Korpusbeispiele einzelner LUs eine sehr stark variierende Auswahl aus den zur Verfügung stehenden Frame-Elementen treffen.

Während der Modellierung wurde also in regelmäßigen Abständen geprüft, ob die entstehende Scenes-Frames-Hierarchie hinsichtlich dieser beiden Kriterien und aus Sicht eines Wörterbuchbenutzers als ausreichend „nützlich“ und „handhabbar“ zu beurteilen ist.[35] War dies nicht der Fall, wurde die vorhandene Struktur modifiziert (gelegentlich auch vollständig verworfen), indem beispielsweise zu umfangreiche Scenes und Frames in mehrere unterteilt oder zu wenig umfangreiche in größere zusammengefasst wurden. Die abstrakten strukturellen Vorgaben des Datenmodells (wie in Abbildung 6 dargestellt) wurden in diesem Prozess allerdings nicht verändert – es ging lediglich darum auszuloten, wie die Freiräume in diesen Vorgaben (etwa: der Abstraktionsgrad einer Scene) optimal zu nutzen sind.

4. Schwierigkeiten bei der Modellierung

Obwohl die im vorigen Abschnitt beschriebene Form der Modellierung sich insgesamt als praktikabel erwiesen hat, haben sich während der Arbeit am Kicktionary dennoch einige Schwächen offenbart. Teilweise liegen diese darin begründet, dass die Methode der Frameanalyse – auch in den Arbeiten Fillmores – in aller Regel lediglich exemplarisch illustriert, nicht aber in Form allgemeingültiger Prinzipien für ein empirisches Arbeiten ausformuliert ist. Die konkrete Vorgehensweise beim Erstellen von Frames und die Beurteilung ihrer „Korrektheit“ bzw. Adäquatheit bleiben daher zu einem großen Teil im Ermessen des Lexikographen, der das empirische Material bearbeitet, und lassen sich kaum durch aus der Theorie abgeleitete Kriterien absichern.

Für die Schwierigkeiten in der Modellierung des Kicktionary war vor allem der Umstand verantwortlich, dass neu hinzukommendes Material (neue lexikalische Einheiten oder neue Beispiele aus dem Korpus) immer wieder die bestehende Scenes- und Frames-Struktur in Frage gestellt hat. Lokale Ergänzungen oder Modifikationen machten somit häufig Änderungen an der globalen Struktur des Wörterbuchs notwendig, deren Auswirkungen sich nur schwer abschätzen ließen, und die in jedem Falle mühsam und zeitaufwendig umzusetzen waren. In einer Diskussion auf der Lexicography-Mailingliste[36] hat Patrick Hanks die Ursache dieses Problems treffend wie folgt beschrieben: „FrameNet proceeds frame by frame, not word by

[35] Mir ist bewusst, dass dies keine im wissenschaftlichen Sinne ausreichend objektivierbaren Begriffe sind. Ob und wie eine solche Objektivierung zu leisten ist (beispielsweise im Rahmen einer wissenschaftlich fundierten Wörterbuchkritik), möchte und kann ich im Rahmen dieses Beitrags nicht diskutieren.

[36] http://tech.groups.yahoo.com/group/lexicographylist/message/3178

word. This may seem a trivial point, but it isn't. Although FrameNet uses empirical data, it does not use an empirical methodology.“

Der oben beschriebene Modellierungsprozess des Kicktionary folgt der hier kritisierten Methode von FrameNet insofern, als er Frames bzw. Scenes vordefiniert und das empirische Material dann in diese Struktur einzufügen versucht. Im Gegensatz zu FrameNet wurde beim Kicktionary aber dennoch versucht, den gesamten relevanten Wortschatz, so wie er sich im Korpus auffinden und belegen lässt, abzudecken. Statt Wörter oder Korpusbeispiele, die nicht in die vorhandene Struktur passen, zunächst oder dauerhaft auszusortieren, musste daher die Struktur selbst fortwährend modifiziert werden, was zu den oben geschilderten Problemen führte.

Um diesen Problemen zu begegnen, scheinen mir zwei Änderungen am Modellierungsprozess bzw. am Modell selbst erforderlich zu sein:

Erstens sollten die Beschreibung einzelner lexikalischer Einheiten und die Annotation zugehöriger Korpusbeispiele nicht von einer übergeordneten Scenes- und Frames-Struktur abhängig sein. Stattdessen sollte sich die Annotation der Argumentstruktur einer lexikalischen Einheit zunächst ausschließlich an der empirischen Analyse der lexikalischen Einheit selbst orientieren. Es ist zu erwarten, dass die auf diese Weise entstehenden Annotationen weniger anfällig für Neu- oder Uminterpretationen sind, einfach weil sie weniger interpretationsabhängige Abstraktion beinhalten. Statt also die Scenes- und Frames-Struktur in einer Top-down-Richtung dem empirisch untersuchten Material aufzuzwängen (bzw. sie in einem schwer zu kontrollierenden Wechselspiel mit der Empirie fortwährend zu modifizieren), könnte und sollte sie in einer Bottom-up-Richtung aus einer Bearbeitung des empirischen Materials auf einer weniger abstrakten Ebene abgeleitet werden. Das Modell selbst würde sich dadurch nicht grundlegend ändern, wohl aber der Modellierungsprozess – die Annotation von Korpusbeispielen würde der Konstruktion der Scenes- und Frames-Hierarchie nämlich dann eindeutig vorausgehen.

Zweitens hat die Arbeit am Kicktionary auch deutlich gemacht, dass Scenes- und Frames-Analysen nicht unbedingt für den gesamten Wortschatz eine gleichermaßen hilfreiche Methode darstellen. Scenes und Frames sind vor allem deshalb nützliche Elemente zur Strukturierung des Wortschatzes, weil sich über sie das (teilweise nichtsprachliche) Wissen über prototypische Handlungsabläufe in Bezug zu sprachlichen Mitteln setzen lässt, mit denen dieses Wissen ausgedrückt werden kann. Große Bereiche der Fußballsprache, wie zum Beispiel die Terminologie für verschiedene Spielerpositionen (*Libero*, *Mittelfeldregisseur*, *Außenstürmer* etc.) oder für Dinge auf dem Spielfeld (*Torpfosten*, *Eckfahne*, *Mittellinie*, *Strafraum* etc.) beschreiben aber gerade keine dynamischen Handlungen, sonder mehr oder weniger statische Objekte. In diesen Bereichen des Wortschatzes sind daher semantische Relationen ein deutlich nützlicheres und auch intuitiver anwendbares Mittel der Wörterbuchstrukturierung als Scenes und Frames. Fillmore (1978, S. 264) selbst erkennt dies auch an:

> I think that semantic theory must reject the suggestion that all meanings need to be described in the same terms. I think, in fact, that semantic domains are going to differ from each other according to the kind of 'definitional base' which is most appropriate to them.

Die Entscheidung, im Modell des Kicktionary eine Zuordnung jeder lexikalischen Einheit zu (genau) einem Frame zu fordern, hat sich vor diesem Hintergrund als nicht sinnvoll erwiesen. Wenn Scenes und Frames ohnehin – wie oben beschrieben – erst entwickelt und definiert werden, nachdem die Beschreibung einzelner lexikalischer Einheiten mit annotierten Korpus-

beispielen abgeschlossen ist, besteht auch keine Notwendigkeit mehr, eine solche Zuordnung zu erzwingen. Die Scenes- und Frames-Hierarchie könnte und sollte sich dann auf denjenigen Bereich des Wortschatzes beschränken, in dem sie sich im Sinne des obigen Zitats als eine „appropriate definitional base“ erweist.

5. Literatur

Date, Chris J. (1995): An Introduction to Database Systems.New York.

Fellbaum, Christiane (Hg.) (1998): WordNet – An Electronic Lexical Database. Boston.

Fillmore, Charles (1978): On the Organization of Semantic Information in the Lexicon. In: Farkas, Donka et al. (Hg.): Papers from the Parasession on the Lexicon, Chicago Linguistic Society, April 14-15, 1978. Reprint in: Fillmore, Charles: Form and Meaning in Language: Volume I, Papers on Semantic Roles. Stanford, S. 261-289.

Fillmore, Charles (1982): Frame semantics. In: Linguistics in the Morning Calm. Selected Papers from SICOL-1981. Edited by the Linguistic Society of Korea. Seoul, S. 111-137.

Fillmore, Charles/Johnson, Christopher R./Petruck, Miriam R. L. (2003): Background to Framenet.In: International Journal of Lexicography 16.3, S. 235-250.

Geeb, Franziskus (2001): leXeML – Vorschlag und Diskussion einer (meta)-lexikographischen Auszeichnungssprache. In: Sprache und Datenverarbeitung 2, S. 27-61.

Kicktionary – The multilingual electronic dictionary of football language: http://www.kicktionary.de/index_de.html (Stand: Oktober 2011).

Ruppenhofer, Josef/Ellsworth, Michael/Petruck, Miriam/Johnson, Chris/Scheffczyck, Jan (2010): FrameNet II: Extended Theory and Practice. Internet: http://framenet.icsi.berkeley.edu/book/book.html (Stand: Oktober 2011).

Schmidt, Thomas (2008): The Kicktionary Revisited.In: Storrer, Angelika/Geyken, Alexander/Siebert, Alexander/Würzner, Kay-Michael (Hg.): Text Resources and Lexical Knowledge. Berlin, S. 239-252.

Schmidt, Thomas (2009): The Kicktionary – A Multilingual Lexical Resource of Football Language.In: Boas, Hans C. (Hg.): Multilingual Framenets in Computational Lexicography. New York, S. 101-134.

Schmidt, Thomas (2010): Der Fußballwortschatz im Kicktionary. In: Der Deutschunterricht 3, S. 17-25.

Stachowiak, Herbert (1973): Allgemeine Modelltheorie. Wien.

Modellierung eines semantischen Wissensnetzes für lexikographische Anwendungen am Beispiel der Duden-Ontologie

Melina Alexa melina.alexa@bi-media.de, Tel. +49 621 3901-330

Die Duden-Ontologie hat mittlerweile eine mehr als 10-jährige Geschichte, von denen ich hier verschiedene Aspekte vorstellen möchte. Zu Beginn stand die Vision alle Duden-Werke in einer zentralen Quelle zu speichern, aus der heraus alle bisherigen und je nach Bedarf auch neue Werke in verschiedenen Formaten und für verschiedene Medien weitgehend automatisch produziert werden können. Darüber hinaus sollten auch sprachtechnologische Produkte diese Quelle nutzen und so von einer permanenten Pflege und kontinuierlichen Überarbeitung und Ergänzung der zentralen Ressource unmittelbar profitieren können. In diesem Papier werde ich zunächst die Motivation und die Ziele erläutern, die uns zu Beginn des Projektes veranlasst haben, uns in dieses Abenteuer zu stürzen. Aus diesen Motiven und Zielen leiteten sich die Anforderungen an die Datenmodellierung ab. Das daraus resultierende Datenmodell werde ich kurz darstellen und anschließend auf die Implementierung eingehen. Zum Schluss gehe ich auf den Einsatz des Wissensnetzes in der Verlagspraxis ein.

1. Motivation und Ziele

Zu Beginn des Projektes im Jahr 2000 wurden die lexikographischen Daten mit einem SGML-basierten Redaktionssystem titelbezogen erstellt, aus dem heraus die Daten für den Drucksatz erzeugt wurden. Obwohl die Inhalte der verschiedenen Wörterbücher sich zum Teil überlappten, mussten die Daten für jedes Werk extra erfasst werden, eine Wiederverwendung war wegen der Werkbezogenheit des Redaktionsprozesses schwierig. Auch die Pflege war ineffizient, da Änderungen, Ergänzungen und Korrekturen der Daten für jedes Werk einzeln nachgearbeitet werden mussten, wie es z.B. während der Rechtschreibreform sehr häufig vorkam. Und neben der Ineffizienz gesellte sich noch die erhöhte Fehleranfälligkeit durch die Doppelarbeit als gravierender Nachteil hinzu.

Daher kam der Wunsch auf, alle Wörterbuchdaten in einer einzigen lexikographischen Ressource zu speichern und zu pflegen, um damit die unterschiedlichen Produkte und Dienstleistungen möglichst flexibel und medienneutral bewerkstelligen zu können. D.h., es sollte möglich sein, eine Neuauflage eines Dudenwörterbuchs aus der zentralen Ressource automatisch in der aktuellen Version aller in Frage kommenden Wörterbuchartikel zu exportieren und das möglichst ohne weitere manuelle Korrekturen an dem zu druckenden Text. Es sollte auch möglich sein, ein komplett neues Wörterbuch zu definieren, d.h. eine Anzahl von Lemmata nach verschiedenen Kriterien festzulegen und ebenso die Inhalte für jeden Eintrag durch Regeln zu definieren, um so den Wörterbuchtext automatisch erzeugen zu lassen. Ob eine Publikation in elektronischer Form, als gewöhnliches Buch oder als eine Kombination aus Buch plus CD angedacht war: dies sollte keinen zusätzlichen Aufwand erfordern. Damals hatte man bei elektronischen Produkten noch eher an CDs oder Online-Angebote gedacht, heute sind es auch E-Books bzw. Apps. Darüber hinaus sollten auch Wörterbuchdaten für sprachtechnologische Software aus der zentralen Ressource exportiert werden können, um auch solche Anwendungen von der permanenten Pflege der lexikographischen Inhalte profitieren zu lassen.

Uns war damals schon klar, dass man ein solch ehrgeiziges Ziel nur erreichen konnte, wenn die Wörterbuchinformationen formal und explizit in einer Datenbank repräsentiert waren. Dazu musste eine effiziente und konsistente Pflege der lexikographischen Inhalte organisiert und gewährleistet werden. Daher wurde ein ambitioniertes Projekt vom Verlag initiiert, an dem die Firma intelligent views GmbH (http://www.i-views.de/web/) und die Fraunhofer Gesellschaft (http://www.ipsi.fraunhofer.de/ipsi/nav/ipsi_f_profile.html) beteiligt waren. Die speziell auf unsere Bedürfnisse hin angepasste Software wurde von intelligent views bereitgestellt und wird bis heute von dort auch gewartet.

Das Ziel dieses Projektes war der Aufbau einer mächtigen Ressource der deutschen Sprache, die sämtliche Informationen und Informationstypen der Dudenwörterbücher beinhaltet und eine Wiederverwendung der Wörterbuchinhalte ohne Informationsverlust für Print- *und* elektronische Anwendungen ermöglicht. Es sollte auch möglich sein, die Ressource durch neue Informationen über die vorhandenen hinaus zu erweitern. Eine redundanzarme Speicherung sowie effiziente Verwaltung, Aktualisierung und Pflege der Wörterbuchinhalte mussten gewährleistet werden. Denn wichtig war von Anfang an, dass alle Duden-(Print-)Wörterbücher mindestens so effizient produziert werden können wie früher aus den verschiedenen titelbezogenen Datenbanken, d.h., es war eine formale und explizite Repräsentation des gesamten lexikographischen Wissens in den Wörterbüchern nötig.

Das Datenmodell muss auch flexibel und erweiterbar sein, um neue Anforderungen, z.B. Repräsentation von weiteren lexikographischen Informationen, erfüllen zu können. Darüber hinaus werden auch mächtige Datenexportmöglichkeiten benötigt, um den vielfältigen Schnittstellen der bei der Produktion nachfolgenden Prozesse genügen zu können.

2. Datenmodell: Duden-Ontologie

Auf Basis all dieser Überlegungen wurde die sogenannte Duden-Ontologie entwickelt. Grundlage unseres Modells ist eine konzeptbasierte Repräsentation, die die Definition semantischer Relationen zwischen den Konzepten ermöglicht. Die Wörterbuchdaten sind mittels einer generischen Hierarchie-Relation klassifiziert, analog zu einer Ontologie.

Eine Ontologie im informatischen Sinne bietet eine formale Methode, Mengen von Individuen zu strukturieren, wobei die Menge der Individuen die Extension eines Begriffs (*Konzept*) ist. Diese Konzepte sind gemäß einer strikten Hierarchie-Relation verbunden. Das ermöglicht die Faktorisierung von gemeinsamen Informationen auf eine abstraktere Ebene. Die wesentlichen Elemente einer Ontologie sind daher typischerweise:

- eine Klassifikation von Konzepten auf Basis der generischen Hierarchierelation (SUB-CONCEPT_OF-Relation) und
- die Unterscheidung zwischen Individuen und Konzepten, bei der ein Individuum zu einem Konzept durch eine INSTANCE_OF-Relation verbunden ist.

Eine Ontologie modelliert eine Bedeutungswelt, in der die Bedeutungen als Konzepte repräsentiert werden. Konkrete Entitäten, wie konkrete Personen, Organisationen, Institutionen, geographische Orte etc., z.B. *Immanuel Kant*, *EU*, *Mannheim*, *IDS*, *Olympische Spiele Athen 2004*, sind darin die Individuen. Entsprechende Konzepte sind *Europäer/-in*, *Philosoph*, *politische Organisation*, *Großstadt*, *wissenschaftliches Institut*, *moderne Olympische Spiele*.

Unsere Idee für die Duden-Ontologie ist, dasselbe Prinzip für die Wörter selbst zu verwenden und eine Ontologie der ‚Welt der Wörter' zu kreieren. Darin werden die Wörter (*Lemmata*) einer Sprache, in unserem Fall der deutschen Sprache, formal als *Individuen* modelliert. Die morphologischen und Grammatikklassen der Sprache, die die Lemmata klassifizieren, sind in diesem Modell *Konzepte.*

Daraus ergibt sich eine Art morphosyntaktische Ontologie über die Welt der Wörter. Diese könnte man als eine weitere Dimension der ersten Ontologie sehen, die die Bedeutungen und die Weltobjekte repräsentiert.

2.1 Term: die Brücke zwischen Lemma und Konzept

Durch diesen Ansatz entstehen also zwei Ontologien: eine ‚normale' Ontologie und eine grammatische Ontologie, die eine strukturiert die Bedeutungen und die andere die Benennungen. Als Brücke zwischen den zwei Ontologien benutzen wir eine Denotationsrelation, um ein Lemma einer Lesart/Bedeutung oder mehreren Bedeutungen zuzuordnen.

Jede Bedeutung eines Lemmas ist eine Rolle, die das Lemma im ‚Sprachspiel' spielt. Jede Rolle ist als einziges Objekt repräsentiert, dieses nennen wir Term. Ein Lemma hat oft mehr als eine Bedeutung, daher können einem Lemma mehrere Terme zugeteilt werden. Jede Bedeutung eines Lemmas ist durch ein einziges Konzeptobjekt repräsentiert.

Auf der anderen Seite können auch einem Konzept mehrere Terme zugeteilt werden, sodass es häufig mit mehr als einem Lemma verbunden ist. Dadurch entsteht die Synonymie-Relation: Zwei Lemmata sind synonym, wenn sie über verschiedene Terme zu dem gleichen Konzept führen.

Man sieht in Abbildung 1 das Top-Konzept der Bedeutungswelt, *Topic*, und das Top-Konzept der morphosyntaktischen Ontologie, *Benennung*. Die Kluft zwischen den Topics und den Lemmata wird durch die Terme überbrückt. Alle Eigenschaften, die für *Topic*, *Benennung* und *Term* im Modell gemeinsam sind, werden zum *BasisObjekt* faktorisiert. Dadurch, dass man Terme als explizite Objekte speichert, können sie auch die Verlinkung zu Anwendungsbeispielen und Zitaten leisten. Auf diese Weise hat man spezifischere Selektionsmöglichkeiten, da so Beispiele für ein Lemma in einer bestimmten Lesart gefiltert werden können.

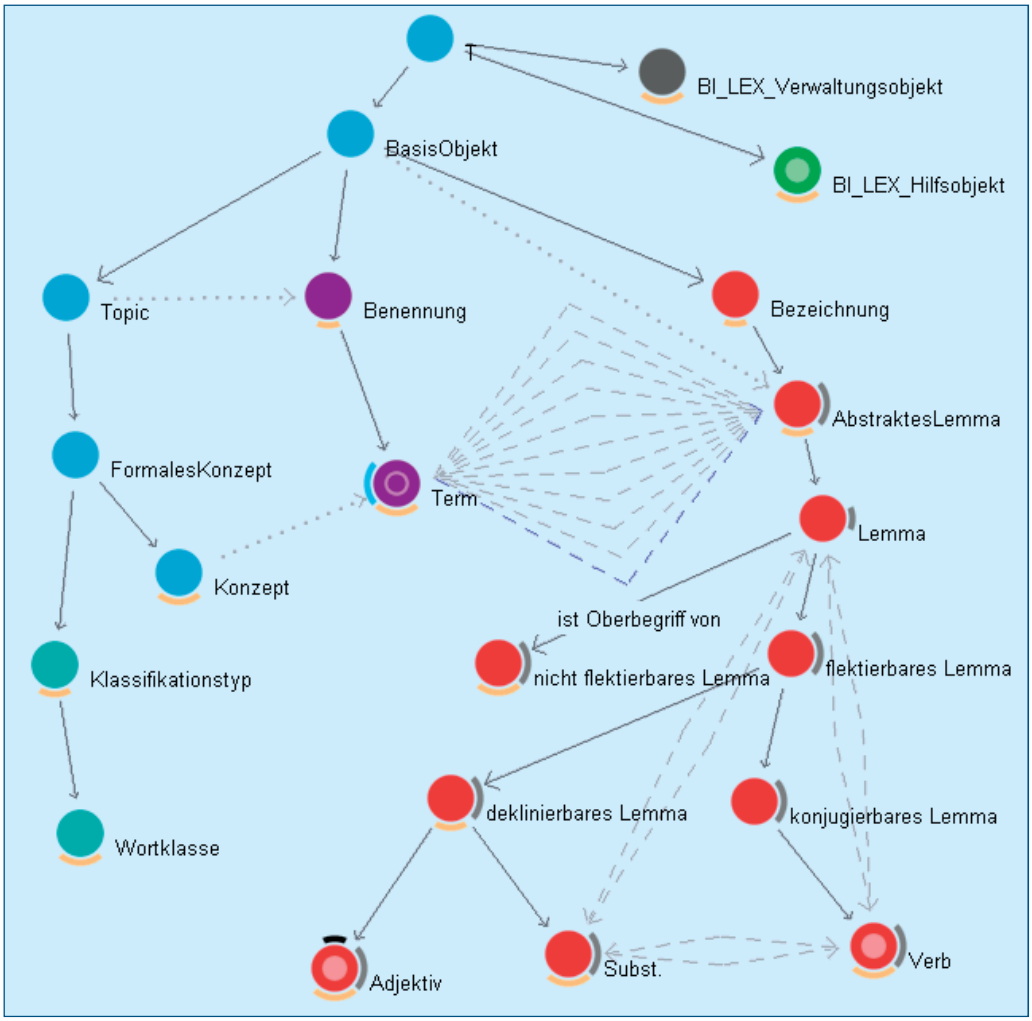


Abb. 1: Top-Level der Duden-Ontologie

Die Wortklassenhierarchie, die die Welt der Wörter gruppiert bzw. klassifiziert, wird in Abbildung 2 gezeigt.

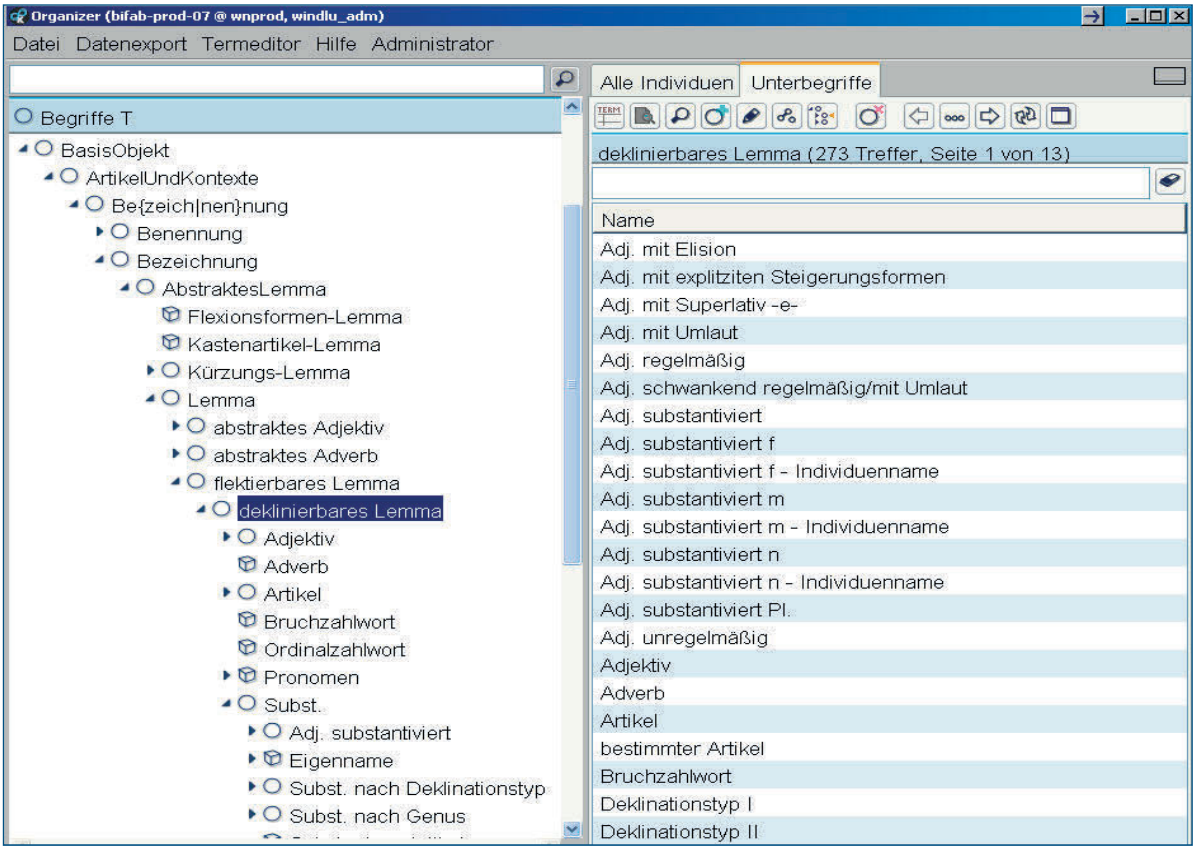


Abb. 2: Wortklassenhierarchie in der Duden-Ontologie

2.2 Implementierung

Die Ontologie ist als Objektnetz repräsentiert. Jedes Konzept steht in Beziehung zu seinen Ober- oder Unterkonzepten. Dadurch können bereits definierte Attribute und Relationen von den allgemeinen zu den spezifischen Konzepten vererbt werden. Des Weiteren gibt es die Möglichkeit, mit multiplen Hierarchien umzugehen; d.h., ein Konzept kann mehrere Oberbegriffe haben, wie am Beispiel der Wortverwendungsklasse in Abbildung 3 zu sehen ist.

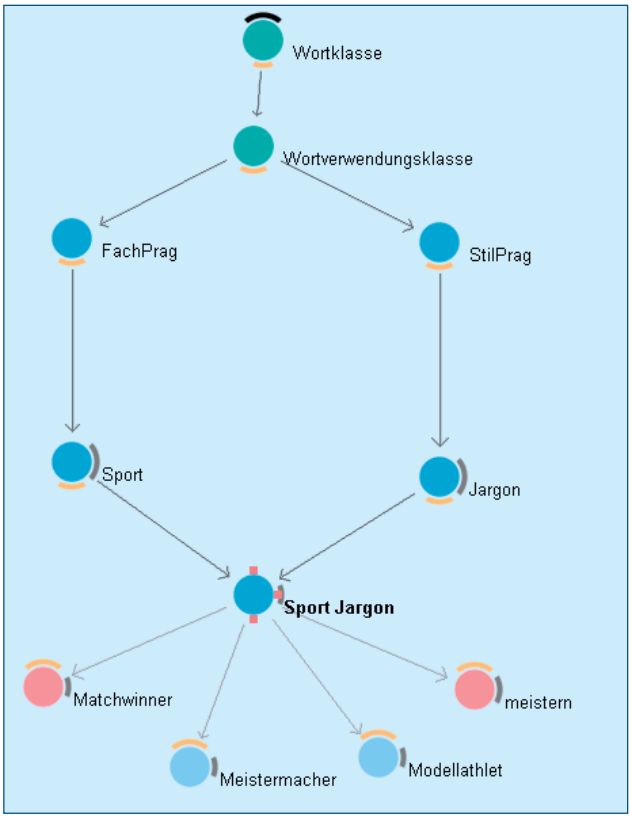


Abb. 3: Beispiel für die Modellierung der Wortverwendungsklasse, Zugehörigkeit der Wörter zu multiplen Hierarchien in der Duden-Ontologie

Das Datenmodell wurde mit der Software K-Infinity von intelligent views implementiert. Die vielfältigen und flexiblen Modellierungsmöglichkeiten sowie die verschiedenen Werkzeuge zur Erzeugung, Bearbeitung, Pflege und Nutzung des Wissensnetzes wurden und werden dabei reichlich genutzt.

3. Import der Wörterbuchdaten ins Wissensnetz

Die ursprünglichen Wörterbuchdaten lagen in SGML vor. Daher bestand der Datenimport aus dem Parsing der SGML-Daten und dem anschließenden Mapping der Elemente und Inhalte in Objekte und Relationen zwischen den Objekten. Typischerweise hat ein Wörterbucheintrag viele verschiedene Informationstypen, u.a. Lemma-Name, Wortklassen-Angaben, Pragmatik- und etymologische Informationen und Anwendungsbeispiele. Für alle diese Informationstypen galt es für den Import entsprechende Mappingregeln zu spezifizieren und dann die resultierenden Objekte und Relationen im Wissensnetz zu generieren.

So wurde jeder Wörterbucheintrag auf ein Lemma-Objekt und die grammatische Information auf eine Grammatikklasse abgebildet. Auch wenn sich das einfach anhört, war dies ein arbeitsintensiver Prozess. Nehmen wir als Beispiel ein Element mit einer spezifischen grammatischen Information: Diese wurde für den Leser des Wörterbuchs geschrieben und nicht für ein Analyse-Programm, um eine systematische Klassifikation in Grammatikklassen zu erreichen. Für solche Fälle mussten wir für den Import in die Duden-Ontologie spezifische Abbildungsprozesse implementieren.

Weitere Herausforderungen waren z.B. die Redewendungen, die wir schließlich als spezifische Lemmatypen behandelt haben und die automatisch während des Imports angelegt wurden. Die Verwendungsbeispiele eines Lemmas mussten einem Termobjekt zugeordnet werden. Da die Terme nur implizit in der Struktur des Wörterbucheintrags vorhanden sind, mussten sie während der Analyse explizit gemacht werden. Auch die Definitionen gehören an die Term-Objekte, d.h., während des Imports musste die jeweilige Bedeutungsvariante ermittelt werden. Wir waren anfangs etwas skeptisch, wie gut das Mapping auf die Terme funktionieren kann. Aber es hat sich herausgestellt, dass die Lesart-Struktur der Wörterbuchartikel gut zu der Termstruktur unseres Modells passt.

3.1 Kompositaanalyse

Leider ist die notwendige Information für ein Mapping nicht immer explizit im Wörterbucheintrag vorhanden, z.B. wird die Grammatikklasse eines Kompositums im Wörterbuch oft nicht angegeben – für den Leser eines Printwörterbuchs ist diese auch nicht zwingend notwendig, denn er oder sie kann sie aus der Angabe der Grammatikklasse für das Grundwort schließen. Für die Zuordnung der Lemmata zu ihren Grammatikklassen aber ist dies vom Modell her zwingend notwendig, z.B. waren im zehnbändigen Duden ca. 50 % der Wörterbucheinträge Komposita, für die eine Lösung gefunden werden musste.

Wir haben eine automatische morphologische Dekomposition durchgeführt, um die Komposita zwischen Grund- und Bestimmungswort zu trennen. Die Ergebnisse dieser Analyse haben wir darüber hinaus genutzt, um diese Relation auch im Netz explizit zu speichern. Wir haben dabei zwei Relationen *hat_Bestimmungswort* und *hat_Grundwort* sowie das Attribut *hat_Fuge* für die Repräsentation der morphologischen Dekomposition definiert.

Die Kompositarelationen für Substantive sind sowohl für Lemmata als auch für Terme, d.h. die einzelnen Lesarten, definiert. Damit Hyperonyme der Komposita verknüpft werden konnten, ist die Termzuordnung unabdingbar gewesen, z.B. ist *Gartenbank* Unterbegriff von *Bank* (als Sitzgelegenheit) und *Investmentbank* ist ein Subkonzept von *Bank* (als Geldinstitut).

3.2 Semiautomatische Extraktion von semantischen Relationen

Ein weiteres Ziel beim Datenimport war es, das Netz mit semantischen Relationen, wie Synonymie, Hyperonymie und Teil-von-Relation zu füllen.

Da jedoch kein explizites Mark-up für solche Informationen in den SGML-Daten vorhanden und eine vollautomatische Akquisition von semantischer Information nicht möglich war, haben wir bestimmte Wörterbucheigenschaften und lexikalische Indikatoren genutzt, um dieses Ziel zu erreichen: Beispielsweise eignen sich Einwortdefinitionen für die Erkennung von Synonymen und auch die Kompositaanalyse selbst eignet sich für die Erkennung von Oberbegriffen, wie eben erläutert. Da die automatischen Verfahren zum großen Teil zwar richtige Ergebnisse liefern, die Extraktion jedoch nicht hundertprozentig korrekte Informationen lieferte, wurden die Ergebnisse nachbearbeitet. Dazu wurden gezielt spezielle Algorithmen und Werkzeuge entwickelt und eingesetzt.

4. Erweiterung des Modells für neue Informationen

Nach den ersten Datenimporten und der ersten Phase unserer Implementierung wurde das Datenmodell um neue Informationen erweitert, z.B. um flektierte Formen (Vollformen), Häufigkeitsklassen und Markierung spezieller Wortschätze, die typischerweise kein Bestandteil von Printwörterbüchern sind, aber z.B. für unser ‚Internetwörterbuch' – Duden online – genutzt werden.

4.1 Vollformen

Im Modell wurden neue Vollformobjekte für Substantive, Verben und Adjektive definiert, die das komplette Flexionsparadigma enthalten. Neu definierte Relationen zwischen den Vollform- und den entsprechenden Lemma-Objekten verknüpfen diese im Wissensnetz.

Somit sind aktuell im Wissensnetz Vollformenobjekte für über 750.000 flektierte Substantive, mehr als 644.000 Verbformen und über 2 Millionen Adjektivformen gespeichert.

4.2 Häufigkeitsklassen

Diese Modellerweiterung betrifft die Information über die Häufigkeitsklasse eines Lemmas. Wünschenswert wäre natürlich, dass die Frequenzanalyse auch für die Terme durchgeführt werden könnte, d.h., sie müsste für jede Lesart die Häufigkeiten ermitteln. Dies könnte nur durch eine semantische Disambiguierung geleistet werden, die jedoch für die deutsche Sprache nicht vorhanden ist. Im Wissensnetz speichern wir daher die Ergebnisse der Frequenzanalyse beim Lemma, somit tragen *Bank*[1] und *Bank*[2] dieselbe Häufigkeitsklasse.

Die Frequenzanalyse wertet dabei das Dudenkorpus (vgl. Münzberg 2011) aus, ein Textkorpus der Gegenwartssprache mit über 2 Milliarden. Wortformen in fünf Textsorten. Die von der Frequenzanalyse ermittelten Angaben zum Wortgebrauch werden in drei Attributen am Lemma gespeichert: *Absolutes Vorkommen* im Korpus, *Korpusrang* nach Häufigkeit und *Frequenzklasse* (selten bis sehr häufig). Daher kann die Häufigkeitsklasse bei den Datenexporten für jedes Lemma problemlos mitgeliefert, als Filterkriterium für die Bearbeitung oder als Angabe für Offline- und Online-Produkte verwendet werden.

4.3 Explizite Markierung von Wortschätzen

Ein anderes Beispiel für eine der aktuellen Modellerweiterungen betrifft die Markierung von Wortschätzen. Konkret umfasst diese die Markierung der Wortschatzzugehörigkeit eines Lemmas mit Hilfe eines Attributes direkt am Lemma. Zurzeit sind der *Grundwortschatz Deutsch als Fremdsprache* und der *Wortschatz des Zertifikats Deutsch* (Goethe-Institut) markiert. Dies ist insbesondere wichtig für Duden online und für neue elektronische und gedruckte Wörterbücher.

5. Steckbrief des Wissensnetzes

Mit dem *Wissensnetz Deutsche Sprache*, wie die Duden-Ontologie auch genannt wird, haben wir also eine umfangreiche linguistische Ressource kreiert. Die dicht geknüpften Daten werden in einer objektorientierten Datenbank gespeichert, die explizit zugreifbares Expertenwissen zu u.a. folgenden Informationstypen enthält:

- Wörter des Deutschen (Stichwörter = Lemmata)
- Rechtschreibung
- Rechtschreibvarianten wie *Dudenempfehlungen* und *Agenturschreibweise*
- Aussprache
- Angaben zur Grammatik
- Bedeutungsangaben
- Synonymie
- Homonymie, Homografie
- Komposita
- Fremdwörter
- Verknüpfung durch semantische Hierarchierelationen
- Flektierte Formen (Vollformen)
- Markierung von Wortschätzen (*Zertifikatswortschatz*, *Duden-Grundwortschatz*)

6. Das Wissensnetz heute in der Praxis

Betrachten wir unseren unmittelbaren Anwendungskontext, die Verlagspraxis, haben wir mit der Duden-Ontologie heute Beeindruckendes erreicht: Importiert wurden bisher die wichtigsten Dudentitel aus unterschiedlichen Wörterbuchreihen, z.B. aus Großwörterbüchern, aus der Reihe „Der kleine Duden“, aus der Schülerdudenreihe oder der Reihe mit den Dudenbänden 1 bis 12. Zu den Wörterbüchern, die im Wissensnetz permanent bearbeitet werden, gehören daher u.a. der Rechtschreibduden und der zehnbändige Duden. Die ins Wissensnetz importierten Titel und Inhalte werden dort integriert gepflegt und aktualisiert. Selbstverständlich werden sie für Neuauflagen und auch für neue und neuartige Produkte verwendet.

Das Wissensnetz ist heute *das* Arbeitswerkzeug der Dudenredaktion und der Duden-Sprachtechnologie, es erlaubt medien- und titelneutrale Bearbeitung von Sprachdaten und titelbezogene Exporte von Sprachdaten für die Buch- und Softwareproduktion. Darüber hinaus ist es die Datenquelle für Duden online und ermöglicht strukturierte Datenexporte zur Überführung in „beliebige“ Formate. Das Wissensnetz bildet dadurch auch die Schnittstelle für unseren Vertrieb für Content- und Datenlizenzgeschäfte und ist nicht zuletzt die Datenquelle für die sprachtechnologischen Produkte aus dem Hause Duden: die „Duden Rechtschreibprüfung“ und den Duden-Thesaurus für Endkunden bzw. die „Duden Proof Factory“ inkl. Thesaurus-Funktionalität für Geschäftskunden.

7. Das Duden-Wissensnetz in einigen Zahlen

Technisch besteht das Wissensnetz heute aus mehr als 1,2 Millionen Individuen-Objekten mit 15,6 Millionen Eigenschaften (Attributen) und 46 Millionen Verknüpfungen untereinander (Relationen).

Inhaltlich umfasst das Wissensnetz insgesamt über 310.000 Lemmata im weiteren Sinne, da auch idiomatische Wendungen in einigen Werken Stichwortstatus haben, darunter über 195.000 Substantive, 22.000 Verben und über 29.000 Adjektive. Des Weiteren gibt es über 1050 verschiedene Wortverwendungsklassen (*umgangssprachlich, veraltend, scherzhaft, derb* etc.) und über 345.000 Definitionen und über 317.000 Ober-/Unterbegriffe.

8. Zusammenfassung und Ausblick

Wir haben mit dem Duden-Wissensnetz eine umfangreiche Ressource über die deutsche Sprache, die eine effiziente Produktion von Print-, Offline- und Online-Wörterbüchern und sprachtechnologischen Anwendungen unterstützt. Die Entscheidungen für die Datenmodellierung und die anschließende Umsetzung wurden in enger Abhängigkeit von unseren Produkten, Anforderungen und Möglichkeiten getroffen: Wir haben ein erweiterbares und integriertes Modell von semantischen und grammatischen Informationen, das die redundanzarme Datenspeicherung unterstützt und unterschiedliche Datensichten mit unterschiedlicher Granularität ermöglicht.

Wir arbeiten kontinuierlich am Ausbau des Wissensnetzes, z.B. durch Ergänzung von weiteren Vollformen und Audio-Daten zur Aussprache. Eine wesentliche Aufgabe für die künftige Arbeit ist die Anreicherung des Wissensnetzes mit zusätzlichen für unsere Produkte und Services wichtigen Informationen, insbesondere für die sprachtechnologischen Anwendungen und für Duden online, dazu zählen z.B. zusätzliches Fachvokabular und die Vertiefung der Vernetzung der semantischen Konzepte mit weiteren Ober- und Unterbegriffen.

9. Literatur

Alexa, Melina/Kreissig, Bernd/Liepert, Martina/Reichenberger, Klaus/Rostek, Lothar/Rautmann, Karin/Scholze-Stubenrecht, Werner/Stoye, Sabine (2002): The Duden Ontology: An Integrated Representation of Lexical and Ontological Information. In: LREC-2002, OntoLex-Workshop 2002: Ontologies and Lexical Knowledge Bases, 27th May 2002, Las Palmas, Canary Islands – Spain.

Münzberg, Franziska (2011): Korpusrecherche in der Dudenredaktion. Ein Werkstattbericht. In: Konopka, Marek, et al. (Hg.): Grammatik und Korpora 2009. Tübingen, S. 181-197. (= CLIP 1).